Semantic-flavored Query Reformulation for Geographic Information Retrieval

Nuno Francisco Pereira Freire Cardoso

Orientadores:
Diana Maria de Sousa Marques Pinto dos Santos
Mário Jorge Costa Gaspar da Silva

Prova de qualificação

Universidade de Lisboa
Faculdade de Ciências

Contents

1 Motivation

One of today's challenges in information retrieval (IR) is the development of new retrieval models that exploit the semantical content of texts to measure the similarity between the users' queries and the collection of documents [Allan et al., 2003]. The main goal of this thesis is to develop more reliable IR approaches based on understanding, rather than in term frequencies and document structure elements which characterise most IR system approaches nowadays [Singhal, 2001].

Query reformulation (QR) is a popular technique to overcome the limitations of classic IR models, and it is widely used among the IR community. In a nutshell, QR adds strongly related terms to an initial query, removes unrelated terms and re-weights the terms according to their importance for the definition of the user's initial information need [Efthimiadis, 1996] (see Figure 1). This process generates queries that are more precise (tackling the vagueness and ambiguity present in most queries with a more strongly related content), and more tolerant to the terminological gap between documents and queries (which is a consequence from the use of a wide vocabulary by different authors on the same subject).

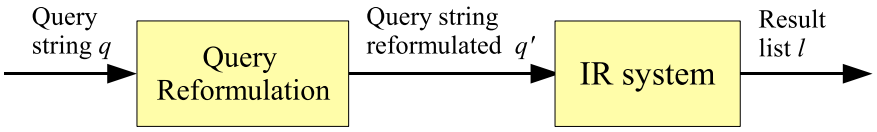

Figure 1: Standard query reformulation.

Most IR systems with QR report improvements on the retrieval results (for example, [Buckley et al., 1995]), as QR is a key component towards the semantic IR desiderata. Yet, most QR approaches are based on term statistics from local feedback (from initial search results) or from global analysis (from the overall collection) [Xu and Croft, 1996], which leads back to the initial problem; in fact, for some imprecise queries, expanded queries may drift from the original topic and even worsen the retrieval results [Mitra et al., 1998].

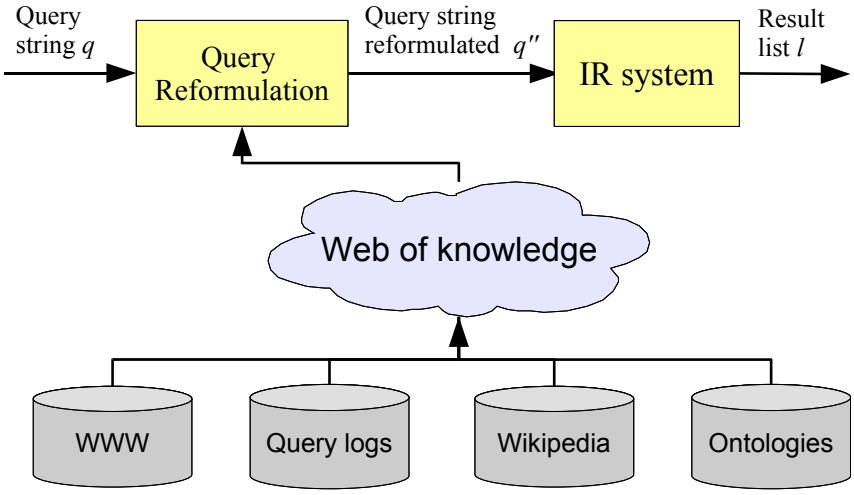

Figure 2: Proposed query reformulation module, assisted by a web of knowledge.

The goal of this thesis is to develop a new automatic QR approach that exploits a web of knowledge to better understand the real information need behind the user who provided the query string, and to reason over such knowledge base to enhance users' queries (see Figure 2). The hypothesis of this thesis is that a semantic-oriented QR module can generate query strings that are more representative of the user's real information need than current state-of-the-art QR modules, and its integration in a modified IR system can achieve better retrieval results than current state-of-the-art IR systems. I define a *web of knowledge* as a semantic network that can be built from several information resources, such as online encyclopedias, query log records and domain specific ontologies.

2 Proposal Context

Geographic information retrieval (GIR) systems feature additional components for capturing and understanding geographic areas of interest (or *geographic scopes*) of both documents and queries, which are used as an additional rank-

ing criteria. QR modules are key components in most GIR systems and thus GIR represents an excellent testbed for the development and evaluation of the proposed semantic-aware QR module.

This proposal contains an appendix document that surveys GIR and related research areas, including a detailed description on the terminology used.

2.1 GIR specific subtasks

GIR differs from IR by facing (at least some of) the following challenging tasks:

- Geographic metadata must be automatically extracted from documents, a task that includes *text mining* and *information extraction* techniques. Queries must also be parsed for placenames and related directions (as in "north of").

- Placenames have several interpretations, regarding their context (for example, "Lisboa" is a term that can be used as a proper name, it can designate several places, refer to an EU treaty or be part of an organization names), whether it points to a place or another thing must be disambiguated (a task known as *named entity recognition*).

- Disambiguated placenames may anyway refer to quite distinct places (for example, "Cuba" is the name of a country, but also the name of a city in Portugal), or different types of places (for example, "Minho" is a placename that can indicate a Portuguese region, or a Portuguese river). The task of finding out which places are referred to by placenames is known as *toponym resolution*.

- Resolved placenames are subsequently mapped to a formal representation of their respective places (a task known as *grounding*). Such places must be represented in a way that GIR systems can perform basic geographic reasoning operations such as computing overlapping, adjacency or distance.

- The geographic metadata generated after grounding must be *indexed* on efficient data structures so that it can be readily available and quickly retrieved.

- Most GIR systems compute automatically the scopes of documents from grounded geographic metadata (an offline process) and scopes of queries from grounded placenames and related directions (an online process). These scopes are used by a *geographic ranking* module to compute geographic similarity between documents and queries.

- Finally, GIR results may be presented to the user with additional geographic information, such as digital maps, so that the user can easily interpret the geographic relevancy of documents and even refine the geographic criteria.

As any IR system, online GIR systems, such as a geographic web search engine, are also concerned with both i) effectiveness, where the top results should be both highly relevant to the user's information need and within its desired geographic scope, and ii) efficiency, where search results should be generated and displayed moments after the query was submitted.

2.2 Projects associated with this thesis

This work follows the successful collaboration of several projects concerned with evaluation, geographic resources and Portuguese – GREASE (I and II), Linguateca and tumba! –, partially or entirely developed at FCUL/XLDB, and within the context of the XLDB node of Linguateca.

GREASE (`http://xldb.di.fc.ul.pt/wiki/Grease`) [Silva et al., 2006] started in 2004 with the purpose of researching methods, algorithms and software architecture for IR systems to provide geographic reasoning over web searches, and it was extended by GREASE II, until 2009.

Linguateca is a distributed network for fostering the computational processing of the Portuguese language (see Santos [2000], Veiga and Santos [2001], Santos [2002], Santos et al. [2004] and Santos [2009] for different snapshots of this

project). Linguateca has been instrumental in fostering evaluation of Portuguese systems and tools, by organizing several evaluation contests for Portuguese and helping disseminate evaluation in the Portuguese-speaking community, namely Morfolimpíadas, HAREM and CLEF. (See Santos [2007], Santos and Cardoso [2007], Peters et al. [2008] and Mota and Santos [2009] for the evaluation effort.)

Tumba! (`http://www.tumba.pt`) [Silva, 2003] is a web search engine specifically designed to archive and provide search services to a Web community formed by those interested in subjects related to Portugal and the Portuguese people. Tumba! has been offered as a public service since November 2002.

GREASE already produced a considerable amount of work, currently bundled in an online geographic web search engine specially built for the Portuguese community, the GeoTumba (`http://local.tumba.pt`). Previous research issues already addressed by GREASE I and Linguateca include:

- the development of a geographic knowledge base that contains detailed information from the administrative and physical domains around the globe, gathered from several public information resources. The geographic knowledge base, GKB, was developed by Chaves et al. [2005]:
- The generation of a geographic ontology from the geographic knowledge base, that represents in detail the Portuguese administrative domain. The ontology, Geo-Net-PT, is publicly available from `http://xldb.di.fc.ul.pt/geonetpt`.
- The identification and disambiguation of placenames in text, and subsequent grounding into geographic concepts from the ontology. Three text mining modules, CaGE [Martins, 2008], Faísca [Cardoso et al., 2008a] and SEI-Geo [Chaves, 2009] were developed for the task.
- The development of a basic QR module as a testbed for experiments with relevance feedback approaches and basic geographic QR. The QR module, QuerCol, was developed by myself [Cardoso et al., 2007].

- The development of a geographic indexing and ranking module to experiment with different approaches to compute geographic relevance between queries and documents, and to combine textual and geographic ranking scores. The resulting module, called Sidra5, was developed by Andrade [2007].

The GREASE II project focuses on:

- the automatic generation of geographic signatures for documents and queries, as a more comprehensive way to represent geographic scopes;
- the automatic query reformulation of geographic queries;
- the research of new approaches to combine geographic and textual rankings;
- a better support for multi-lingual documents;
- the addition of physical geography of the world in the knowledge base;
- the development of faceted interfaces for GIR.

Evaluation is an essential task for leveraging our current GIR model against other models developed by other GIR researchers, and provides valuable input on the advantages and failure points for the GIR components. My work will therefore include participation and co-organization of the forthcoming GikiCLEF (`http://www.linguateca.pt/gikiclef/`), a new CLEF evaluation track (under the QA@CLEF umbrella) that evaluates systems on the task on finding Wikipedia entries / documents that answer a particular information need which requires geographical reasoning of some sort.

3 My view on QR and GIR

Information Retrieval (IR) is a research area with over 50 years of activity, devoted to the problem of searching relevant information over large collections of

documents. A typical IR system receives a list of terms as input (a *query string*), and then searches for documents that match the given terms, returning a list of results ordered by relevance.

While this approach suffices for some types of search, it often fails to retrieve relevant documents for queries that may be vaguely expressed, or for queries that have a more complex underlying information need. For instance, the query string "restaurante Santa Bárbara" is not clear about what the user wants: is it i) pages about restaurants within a certain place called Santa Bárbara, or ii) the home page of a given restaurant called "Santa Bárbara"? Either way, a classic IR system will return documents ordered by a ranking algorithm based on term statistics, and the user may have a hard time filtering out the documents that are indeed relevant to him or her.

To properly handle such queries, the ideal IR system should be able to reason that "restaurante Santa Bárbara" is an ambiguous query (where Santa Bárbara might be a placename or a restaurant name), detect the true meaning of the query (for example, the search history of the user might suggest that he is interested on visiting the city of Santa Bárbara in Minas Gerais, Brazil), and rank the results according to that meaning (for example, presenting the results according to the geographic proximity to the city of Santa Bárbara, and displaying a digital map with pinpoint locations).

Note that, in my view, the QR task is not limited to the act of re-writing query strings – this is just the final outcome of the overall QR module. QR includes all kinds of approaches made, from reading the initial query to writing a final string, that are exclusively dedicated to capture the key elements of the user's needs, and to rewrite such needs in a way that the underlying retrieval engine can efficiently search for documents that do match the user's expectations.

While IR systems typically use QR modules to generate more query terms and improve retrieval performance by narrowing the terminological gap between documents and queries, a GIR system needs QR modules to interpret geographically-

related queries and present it to the document ranking module, which now computes ranking scores for both text similarity and geographic relevance.

My perspective is that a GIR system, in order to successfully perform its challenging task, should:

1. Employ NLP-based techniques for query reformulation approaches, to better capture the real intentions of users beneath simple user strings, and thus better prepare a final, unambiguously query string that renders such message to the retrieval engine in an appropriate way.

2. Exploit all kinds of information resources that can provide any kind of additional knowledge that helps understanding the user queries. As user needs vary from geographically-related topics to navigational queries [Aires, 2005], such information resources should not just based on factual information, but also include information about user search trends and the world-wide-web structure.

While the focus of this thesis is to develop a new QR module based on semantic approaches, the work of this thesis will also include the development and improvement of other GIR modules to implement the new approach, thus implying the development of a new GIR prototype system.

A major challenge for the semantic vision of an IR system is the automatic access to a web of knowledge that spans all potentially searchable topics in a format that can easily understood by systems [Berners-Lee et al., 2001]. The present proposal exploits four information sources consisting a web of knowledge to assist the GIR system: i) ontologies, ii) world wide web, iii) Wikipedia, and iv) query logs (see Figure 3). I now introduce each information resource and give examples on how to exploit each resource to gather information about the term "Lisbon" (to keep the example simple, assume that Lisbon is already resolved to its geographic concept of a capital city).

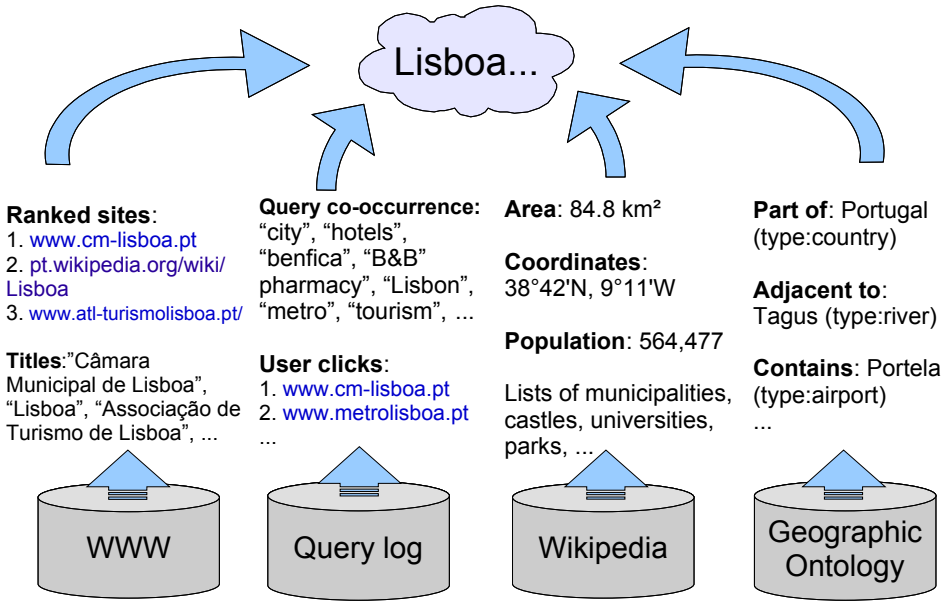

Figure 3: Using a web of knowledge to gather information about "Lisbon".

3.1 Ontology

Ontologies are a way to describe a domain, encoding concepts and relationships between concepts in a machine-friendly format (e.g., XML/RDF). Regarding the geographic domain, geographic ontologies such as Geo-Net-PT, the first geographic ontology of Portugal, are crucial to provide a basis for geographic knowledge and reasoning for GIR systems. For instance, a geographic ontology has information on how the geographic concept "Lisbon" relates to other neighbour geographic concepts. As such, the QR module can infer which geographic concepts are related to Lisbon, such as adjacent rivers (Tagus), and use this information to reason over complex queries as in "canoeing activities in Lisbon".

3.2 World-wide web

The world-wide web is probably the largest information resource ever built. While its content can be mined to extract knowledge [Etzioni et al., 2005], my hypothesis is that an online encyclopedia such as Wikipedia is more suitable for such purpose. On the other hand, a snapshot of the web, properly downloaded and mined, can provide global information about the morphology of the web (for instance, the most popular sites) and search feedback (URLs, titles and surrogates of top-ranked documents for a given query). For instance, the query "Lisboa Editora" is likely to return the homepage of Lisboa Editora (`http://www.lisboaeditora.pt/`) and related documents about this company, which may suggest that such query has a navigational purpose, and that "Lisboa" is just part of the company's name, not a geographic criteria. Also, the world-wide-web can provide the information to classify named entities that are not have a Wikipedia article, such as "Lisboa Editora".

3.3 Wikipedia

The online encyclopedia Wikipedia (`www.wikipedia.org`) has a considerable amount of documents describing concepts and entities in an accurate and structured way, thanks to thousands of anonymous contributors and reviewers. Wikipedia is now being widely used as a resource for bootstrapping Semantic Web systems such as DBpedia [Auer et al., 2007]. For instance, the Wikipedia article about Lisbon contains additional properties such as its geographic coordinates, its area and associated population, and encompasses several topics about the city, such as historical events or tourism information, which can be of use for QR when such subjects were requested by the user.

3.4 Query logs

Web server logs are a valuable resource for QR, as they represent past users' relevance feedback on query/document pairs, among other things. Manual query re-

	Ontologies	WWW	Wikipedia	Query Logs
Accessibility	-	++	++	++
Information credibility	++	-	+	-
Subject diversity	-	++	++	+
Domain specificity	++	-	+	- -
Machine-friendly format	++	-	+	-
Information freshness	-	+	++	-
User content	- -	-	- -	++

Table 1: Characteristics of the information sources.

formulations of users can also be explored to learn QR techniques, or user clicks on same documents can be used to cluster query terms. For instance, term co-occurrence statistics from past users' queries with the term "Lisbon" (as in "Lisbon zoo" and "Lisbon beaches") may represent different information needs as reflected in the different documents selected by the users. Conversely, searches like "Lisbon Metro" and "Lisbon Subway" might be followed by page clicks on the same result (`http://www.metrolisboa.pt`), indicating that "Subway" and "Metro" are two strongly related terms.

3.5 Characteristics of the information sources

Table 1 summarises the characteristics of each resource type and their contribution for the web of knowledge, which will be further detailed in this proposal. The access to Wikipedia snapshots is free to any user, while WWW snapshots are more difficult to get hold of access for non-research purposes. Query logs are hard to access for the general public, due to privacy issues, but for the present research work the query logs of tumba! are available. Ontologies have a high level of credibility, as they are carefully reviewed and validated. Wikipedia and its vast community that updates and verifies its contents (either by manually editing pages or by executing and supervising Wikipedia maintenance tools) make it also a highly credible source of information. The WWW does not have any

kind of restrictions on the published information, hence its credibility is indirectly estimated by the authority of the host site, for example.

Ontologies are a typical choice for an accurate representation of a given domain, and as such their scopes are normally confined to that domain. The WWW and the query logs are quite the opposite, as they span a large variety of subjects. Wikipedia represents an interesting compromise, allowing an hierarchical organization of subjects through a range of categories, without restraining the subject diversity (provided, of course, that subjects lie within certain ethical patterns and have some pertinence to the common interest).

Regarding format, ontologies are the most machine-friendly resource, as its structured format and underlying languages, XML/RDF/OWL, are the de facto format for knowledge representation for machines. The Wikipedia structure is also suitable to be automatically mined, while the WWW still poses data cleaning challenges. Query logs, although they have a tab-separated field format, have no structure at all regarding resource description of user interactions. Tumba! logs include several additional information such as session identifiers and clicked navigational links, which can be mined to extract information regarding users' search habits, for instance analysing the average session time, or aggregate several queries related to the same search (see the preliminary work of Seco and Cardoso [2006]).

Wikipedia generates periodic snapshots of its contents in XML and SQL formats, and as such it has a high degree of information freshness. Although the WWW is theoretically always 'fresh', the crawling procedure takes a considerable amount of time and thus some documents may not be precisely up to date [Gruhl et al., 2004]. On the other hand, ontologies have the lowest refreshing ratio, because they require human-expert revision and validation of new data.

Finally, the most valuable characteristic of the query logs is that they have information regarding users' searches and topics of interest, while the other resources do not contain such user feedback data.

4 Objectives and Contributions

The main objective of this thesis is to develop a new automatic QR module based on semantic approaches that effectively understands the real information need behind user queries, so that its integration in a GIR system leads to a significant improvement of the retrieval results for geographically-related queries.

The objective can be divided as:

- Formulate geographic QR as a special kind of QR specialised in the geographic domain, by identifying its main challenges, requirements and lines of research.
- Exploit a web of knowledge built from several resources, such as online-encyclopedias or specific ontologies, for assisting a specially crafted semantic-oriented QR module.
- Evaluate the suitability of the information resources and NLP-based QR techniques to the overall GIR task.

The objectives will be assessed by:

- Measuring the impact of QR on the overall performance of a GIR system. The performance gain of the geographic QR module with different configurations and extracted geographic knowledge will be measured.
- Characterising the differences between semantic approaches for QR and state-of-the-art QR statistical approaches. The two paradigms will be compared by measuring the performance gain for a common set of query topics and a common document collection.
- Evaluating the usefulness of the QR module. The QR module will be deployed into the online web GIR system developed within the GREASE project, GeoTumba, and user satisfaction studies will be performed and analysed.

The expected results of this work are:

- A semantic-flavored query reformulation module specially crafted for Portuguese, called RENOIR, which uses NLP-based approaches and exploits a web of knowledge to reformulate query strings that have a direct impact on the overall improvement of retrieval performance for geographic queries. RENOIR will handle users' queries in an efficient and effective way, as a component of the online geographic web search engine GeoTumba.

- A stable version of REMBRANDT, a named entity recognition and entity relation detection module based also on semantic approaches and a web of knowledge for its core operation. REMBRANDT shall be capable of capturing most geographic evidence from placenames and other related entities in Portuguese texts, and provide this additional knowledge for QR. REMBRANDT differs from the other state-of-the-art NER systems by using natural language processing techniques on a web of knowledge derived from Wikipedia (or similar resources).

The modules will become publicly available, under the GPL license and with source code included.

The relevant scientific contributions to research in query reformulation, information extraction, text mining and geographic information retrieval will be published in major conferences and journals of these research areas.

5 Work Plan

5.1 Preliminary work on GIR

The first GIR approach taken by GREASE I postulated that both queries and documents had their geographic scope represented by a single encompassing geographic reference to a geographic concept contained in an ontology [Martins, 2008]. Geographic relevance of a document to a query was computed through

a set of heuristics that measure distance and overlapping areas, population count and ontological relationships.

Although that approach had some advantages (as shown by evaluation results reported in Martins et al. [2007]), its limitations led to the following modifications [Cardoso et al., 2008a, 2009]:

Shallow text mining The CaGE text mining tool often failed to capture important geographic metadata from documents, thus having a poor recall on documents with a geographic scope assessed by human annotators. REMBRANDT was developed precisely to provide a more robust text mining tool for geographic metadata.

Implicit geographic evidence Placenames represent a relevant source of geographic evidence of the scope of the document (and called explicit external evidence), but there are other entities such as postal codes, organizations (through their headquarters or branches) or events (through the place where they occurred), that also have a strong geographic connotation, although they are not placenames (called implicit geographic evidence). REMBRANDT recognizes such entities and associates them to their underlying geographic places. A first experiment was made to evaluate how GIR retrieval improves with such implicit geographic data [Cardoso et al., 2008b], albeit with inconclusive results.

Handling subject and geographic criteria on queries Queries were treated as <*what*, *spatial relationship*, *where*> triplets, where each part was handled differently by the GIR components. As this query parsing practice has not yet proven its merits [Cardoso and Santos, 2008], query terms are now handled in a non-segregational way. More work on characterising queries has yet to be done.

Geographic scopes represented as geographic signatures Instead of being described by a single geographic reference, geographic scopes are now de-

scribed by *geographic signatures*, that is, a document surrogate that contains only named entities related to geographic places, weighted according to their relevance to the scope [Cardoso et al., 2008a].

Computing document ranking scores The initial heuristic approaches for geographic relevance were replaced by a BM25 term weighting scheme [Robertson et al., 1992] adapted to weigh both term and geographic indexes [Cardoso et al., 2009]. A comparison between the two geographic ranking approaches is on the work plan of this thesis.

Smoothen geographic QR The addition of new terms to the initial query may sometimes lead to a drift from the initial topics, with a negative impact on the retrieval results. Query reformulation now includes a term re-weighting step, providing a way to give different importance scores for expanded terms and placenames, according to their relevance for the topics [Cardoso et al., 2009].

5.2 Software developed

5.2.1 QuerCol

QuerCol was initially deployed in 2004 by myself, as a query reformulation platform to provide basic query expansion capabilities for a prototype IR system to participate in CLEF for the first time [Cardoso et al., 2005]. With the GeoCLEF task in 2005, QuerCol was extended to provide basic geographic query expansion capabilities, using a geographic ontology as a source for related names [Cardoso et al., 2006].

Later in 2006, QuerCol was improved to experiment with different query reformulation expansion strategies according to the features, feature types, and spatial relationships present in the queries [Cardoso and Silva, 2007]. Given the example query string “hotéis nas ilhas portuguesas”, QuerCol selected its geographic QR strategy by searching all geographic concepts in the ontology that were both of

type “island” and related to the geographic concept “Portugal” by a *part-of* relationship. In the end, the final query string should contain all names of Portuguese islands as the geographic scope of interest, which is a more comprehensive and robust query to use in the retrieval process. Additionally, all Portuguese islands are grounded into geographic concepts in the ontology, which is important for the computation of the geographic similarity between scopes of documents and queries.

5.2.2 REMBRANDT

REMBRANDT is a language-dependent named-entity recognition (NER) system that uses Wikipedia as a raw knowledge resource, and explores the Wikipedia document structure to classify all kinds of named entities in the text. REMBRANDT's development started in 2008 by myself (more information regarding its strategy and implementation details can be found in Cardoso [2009]).

By using Wikipedia, REMBRANDT obtains additional knowledge on every named entity that can be useful for understanding the context, detecting relationships with other named entities, and use this information to contextualize and classify surrounding named entities in the text. REMBRANDT currently classifies named entities using the 9 main categories and 47 sub-categories as defined in the Second HAREM [Santos et al., 2008]. The main categories are: `PERSON`, `ORGANIZATION`, `PLACE`, `DATETIME`, `VALUE`, `ABSTRACTION`, `EVENT`, `THING` and `WORKS`. REMBRANDT can handle vagueness in named entities, by assigning more than one category or sub-category to the named entity.

REMBRANDT participated in the Second HAREM, where it obtained an F-measure of 0.567 for the full NER task and ranked as the 2nd best system out of 10, and ranking first out of 8 systems for the `PLACE` only scenario with an F-measure of 0.625. Regarding the ReRelEM task, which evaluated entity relation detection, REMBRANDT achieved the best results for the `PLACE` only scenario (F-measure of 0.727), showing good capabilities for detecting relations between placenames.

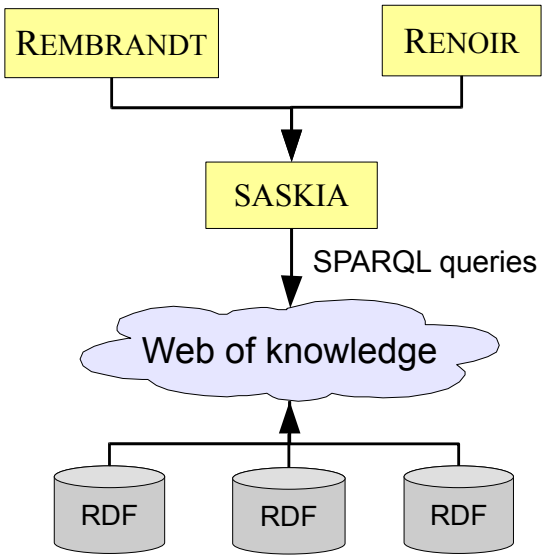

Figure 4: The SASKIA module, serving information of the web of knowledge to other applications.

5.2.3 RENOIR

RENOIR started in 2008 as a proof-of-concept for new GIR approaches suited for the GikiP pilot task [Santos et al., 2009], which predated GikiCLEF, and is currently being re-implemented to become a fully-automated QR system which can use information from the web of knowledge to better redefine the concepts on the original query string.

5.2.4 SASKIA

SASKIA started as a component of REMBRANDT that provided named entity classifications based solely on Wikipedia information, along with an API to the Wikipedia data. Now, SASKIA will take care of providing a common query interface and API service for all information resources, to all semantic applications that need access to the web of knowledge.

Figure 4 shows how the SASKIA assists semantic applications by serving the the information resources in a XML/RDF format that can be queried by SPARQL statements [Prud'hommeaux and Seaborne, 2008]. This allows REMBRANDT and RENOIR to be more focused on reasoning strategies rather than on accessing and managing information.

5.3 Work planned

To achieve the proposed scientific objectives, the work planned for this thesis envisions:

- the improvement of the software components REMBRANDT and RENOIR to achieve the best performance possible, and the selection and modification of an indexing and ranking module that best suits the GIR module. For this, I will evaluate software packages that implement different approaches on IR retrieval and ranking, namely MG4J [Boldi and Vigna, 2005] (based on the BM25 term weighting scheme), LM-Lucene[1] (a Lucene[2] package with a language model extension) and TERRIER[3] (a package that implements the divergence from randomness model).

- RENOIR, REMBRANDT and the indexing & ranking module will be evaluated separately, to measure the fitness of each implemented approach, and detect failure points and system bottlenecks. This will be achieved by participating in specific evaluation contests for each component, such as the REMBRANDT module and the HAREM evaluation contest [Santos and Cardoso, 2007]. This ensures that the modules have the best performing versions of the software, and their overall results is a direct consequence of the approaches proposed on this thesis.

[1] `http://ilps.science.uva.nl/resources/lm-lucene`
[2] `lucene.apache.org`
[3] `ir.dcs.gla.ac.uk/terrier/`

- The modules will be assembled into a GIR prototype that will be compared against other state-of-the-art IR and GIR systems in controlled evaluation environments given by past and future editions of international evaluation conferences, such as GeoCLEF and GikiCLEF. This ensures that the work is repeatedly evaluated in an unbiased environment, and its viability measured against alternative approaches for achieving a same task.

- GIR system evaluation will also include a thorough analysis of the retrieval results, namely a detailed study on the impact of the types of geographic queries on the retrieval performance (as noted by Santos and Chaves [2006]), and a linguistic analysis over relevant documents that failed to be retrieved.

5.4 Calendar

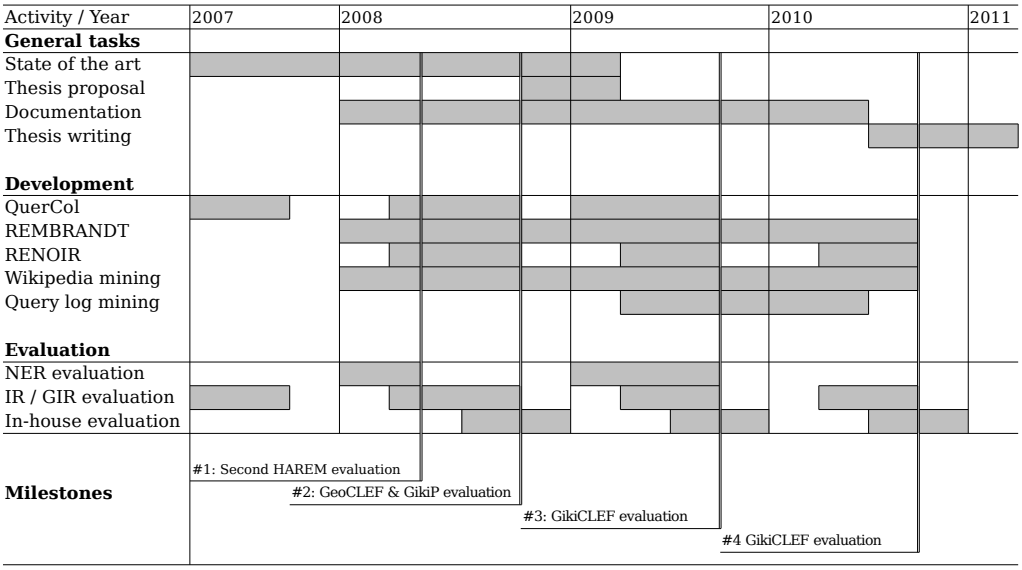

Figure 5: Calendar for the PhD thesis.

The expected calendar for this work is presented in Figure 5. It contains four milestones that match the past HAREM, GeoCLEF and GikiP evaluations and the forthcoming GikiCLEF evaluations (the GikiCLEF task is confirmed only for 2009; future HAREM evaluations are not yet confirmed), where the new versions of the modules will be tested. Work is therefore divided in four parts, defined by those milestones:

Second HAREM evaluation, where a stable version of REMBRANDT was developed and evaluated over a collection of Portuguese documents.

GeoCLEF and GikiP evaluations of 2008, where REMBRANDT was used to annotate a document collection with more than 200,000 documents, an initial version of RENOIR was developed to have first contact with the new challenges posed by GikiP, and where significant modifications were made to the MG4J indexing and ranking module. The overall results of the GIR system were encouraging [Cardoso et al., 2009] and are now the basis for the work proposed on this thesis.

GikiCLEF evaluation of 2009, which will deeply evaluate the improved RENOIR module on its ability to use the Wikipedia information and the DBpedia resources, and on the preliminary semantic QR approaches.

CLEF evaluation of 2010, using a final version of REMBRANDT, capable of achieving state-of-the-art NER performances for both Portuguese and English texts. RENOIR will make now extensive use of the web of knowledge, and will produce complex query strings. The performance gain with these successive versions of the modules will be measured, to assess the retrieval improvements achieved with these new approaches.

Acknowledgement

This work is supported by Fundação para a Ciência e Tecnologia under the scholarship grant SFRH/BD/29817/2006, projects GREASE (POSI/SRI/47071/2002) and GREASE II (PTDC/EIA/73614/2006), and co-supported by POSI under the project Linguateca (POSC/339/1.3/C/NAC).

References

Rachel Aires. *Uso de marcadores estilísticos para a busca na Web em português*. PhD thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Agosto 2005. in Portuguese.

James Allan, Jay Aslam, Nicholas Belkin, Chris Buckley, Jamie Callan, Bruce Croft, Sue Dumais, Norbert Fuhr, Donna Harman, David J. Harper, Djoerd Hiemstra, Thomas Hofmann, Eduard Hovy, Wessel Kraaij, John Lafferty, Victor Lavrenko, David Lewis, Liz Liddy, R. Manmatha, Andrew McCallum, Jay Ponte, John Prager, Dragomir Radev, Philip Resnik, Stephen Robertson, Roni Rosenfeld, Salim Roukos, Mark Sanderson, Rich Schwartz, Amit Singhal, Alan Smeaton, Howard Turtle, Ellen Voorhees, Ralph Weischedel, Jinxi Xu, and ChengXiang Zhai. Challenges in Information Retrieval and Language Modeling: Report of a Workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts, Amherst, September 2002. *SIGIR Forum*, pages 31–47, 2003.

Leonardo Andrade. Processing Geographic Queries and Architectural Experiments with the Tumba! Search Engine. Master's thesis, University of Lisbon, Faculty of Sciences, December 2007.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A Nucleus for a Web of Open Data. In Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11–15, 2007, Proceedings*, number 4825 in LNCS, pages 722–735. Springer, 2007.

Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, May 2001.

Paolo Boldi and Sebastiano Vigna. MG4J at TREC 2005. In *Proceedings of the 14th Text REtrieval Conference (TREC'2005)*. NIST Special Publication SP 500-266, 2005. `http://mg4j.dsi.unimi.it`.

Chris Buckley, Gerald Salton, James Allan, and Amit Singhal. Automatic Query Expansion Using SMART: TREC 3. In *Proceedings of The 3rd Text REtrieval Conference (TREC-3)*, pages 69–80, 1995.

Nuno Cardoso. REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e Análise Detalhada do Texto. In Cristina Mota and Diana Santos, editors, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2009.

Nuno Cardoso and Diana Santos. To Separate or not to Separate: Reflections About GIR Practice. In *1st Workshop on Novel Methodologies for Evaluation in Information Retrieval (NMEIR'08)*, Glasgow, UK, March 30 2008.

Nuno Cardoso and Mário J. Silva. Query Expansion through Geographical Feature Types. In *Proceedings of the 4th Workshop on Geographic Information Retrieval (GIR'07)*, Lisbon, Portugal, November 9 2007. ACM.

Nuno Cardoso, Mário J. Silva, and Miguel Costa. The XLDB Group at CLEF 2004. In Carol Peters, Paul Clough, Julio Gonzalo, Gareth Jones, Michael Kluck, and Bernardo Magnini, editors, *Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004*, volume 3491 of *LNCS*, pages 245–252. Springer, 2005.

Nuno Cardoso, Bruno Martins, Leonardo Andrade, Marcirio Chaves, and Mário J. Silva. The XLDB Group at GeoCLEF 2005. In Carol Peters, Frederic Gey, Julio Gonzalo, Henning Müeller, Gareth J.F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke, editors, *Acessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers*, volume 4022 of *LNCS*, pages 997–1006. Springer, 2006.

Nuno Cardoso, Mario J. Silva, and Bruno Martins. The University of Lisbon at CLEF 2006 Ad-Hoc Task. In Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, and Maximilian Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006, Alicante, Spain, September, 2006. Revised Selected papers*, volume 4730 of *LNCS*, pages 51–56. Springer, Berlin, 2007.

Nuno Cardoso, David Cruz, Marcirio Chaves, and Mário J. Silva. Using Geographic Signatures as Query and Document Scopes in Geographic IR. In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, Anselmo Peñas, Vivian Petras, and Diana Santos, editors, *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, volume 5152 of *LNCS*, pages 802–810. Springer, 2008a.

Nuno Cardoso, Mário J. Silva, and Diana Santos. Handling Implicit Geographic Evidence for Geographic IR. In James G. Shanahan, Sihem Amer-Yahia, Ioana Manolescu, Yi Zhang, David A. Evans, Aleksander Kolcz, Key-Sun Choi, and Abdur Chowdhury, editors, *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM 2008)*, pages 1383–1384, Napa Valley, CA, USA, October 26-30 2008b. ACM.

Nuno Cardoso, Patrícia Sousa, and Mário J. Silva. Experiments with Geographic Evidence Extracted from Documents. In Carol Peters, Tomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F.Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, and Viviane Petras, editors, *Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*. Springer, 2009.

Marcirio Chaves. Geo-ontologias e padrões para reconhecimento de locais em textos: a participação do SEI-Geo no segundo HAREM. In Cristina Mota and Diana Santos, editors, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: Actas do Encontro do Segundo HAREM*. Linguateca, 2009.

Marcirio Chaves, Mário J. Silva, and Bruno Martins. A Geographic Knowledge Base for Semantic Web Applications. In Carlos A. Heuser, editor, *20 Simpósio Brasileiro de Bancos de Dados (SBBD'2005)*, pages 40–54, Uberlândia, MG, Brazil, October 3-7 2005.

Efthimis N. Efthimiadis. Query Expansion. *Annual Review of Information Systems and Technology*, 31:121–187, 1996.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Unsupervised Named-Entity Extraction from the Web: An Experimental Study. *Artificial Intelligence*, 165(1):91–134, 2005.

Daniel Gruhl, Laurent Chavet, David Gibson, Jörg Meyer, Pradhan Pattanayak, Andrew Tomkins, and Jason Y. Zien. How to Build a WebFountain: An Architecture for Very Large-Scale Text Analytics. *IBM Systems Journal*, 43(1): 64–77, 2004.

Bruno Martins. *Geographically Aware Web Text Mining*. PhD thesis, University of Lisbon, Faculty of Sciences, August 2008.

Bruno Martins, Nuno Cardoso, Marcirio Chaves, Leon ardo Andrade, and Mário J. Silva. The University of Lisbon at GeoCLEF 2006. In Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, and Maximilian Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of*

the Cross-Language Evaluation Forum, CLEF 2006. Alicante, Spain, September, 2006. Revised Selected papers, volume 4730 of *LNCS*, pages 986–994. Springer, Setember 2007.

Mandar Mitra, Amit Singhal, and Chris Buckley. Improving Automatic Query Expansion. In *Proceedings of the 21st Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'98)*, pages 206–214, Melbourne, Australia, August 24-28 1998. ACM Press.

Cristina Mota and Diana Santos. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2009.

Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, Anselmo Peñas, Vivien Petras, and Diana Santos. *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, volume 5251 of *LNCS*. Springer, 2008.

Eric Prud'hommeaux and Andy Seaborne. SPARQL Query Language for RDF. Technical report, W3C, January 2008. URL `http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/`.

Stephen E Robertson, Steven G. Walker, Micheline Hancock-Beaulieu, Aarron Gull, and Marianna Lau. Okapi at TREC. In *Proceedings of the 1st Text REtrieval Conference (TREC'92)*, pages 21–30. National Institute of Standards and Technology (NIST), 1992. Special Publication 500-207.

Diana Santos. O projecto Processamento Computacional do Português: Balanço e perspectivas. In Maria das Graças Volpe Nunes, editor, *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000)*, pages 105–113, São Paulo, SP, Brazil, November 19-22 2000. ICMC/USP.

Diana Santos. Um centro de recursos para o processamento computacional do português. *DataGramaZero - Revista de Ciência da Informação*, 3(1), February 2002. http://www.dgz.org.br/fev02/Art_02.htm.

Diana Santos. *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. IST Press, Lisboa, Portugal, 2007.

Diana Santos. Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva. *Linguamática*, 2009.

Diana Santos and Nuno Cardoso. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, 2007.

Diana Santos and Marcirio Chaves. The Place of Place in Geographical IR. In Ross Purves and Chris Jones, editors, *Proceedings of the 3rd ACM Workshop On Geographic Information Retrieval, GIR 2006, Seattle, WA, USA, August 10, 2006*. Department of Geography, University of Zurich, 2006.

Diana Santos, Alberto Simões, Ana Frankenberg-Garcia, Ana Pinto, Anabela Barreiro, Belinda Maia, Cristina Mota, Débora Oliveira, Eckhard Bick, Elisabete Ranchhod, José João Dias de Almeida, Luís Cabral, Luís Costa, Luís Sarmento, Marcirio Chaves, Nuno Cardoso, Paulo Rocha, Rachel Aires, Rosário Silva, Rui Vilela, and Susana Afonso. Linguateca: um centro de recursos distribuído para o processamento computacional da língua portuguesa. In Guillermo De Ita Luna, Olac Fuentes Chávez, and Mauricio Osorio Galindo, editors, *Proceedings of the international workshop "Taller de Herramientas y Recursos Linguísticos para el Espanõl y el Portugués" and IX Iberoamerican Conference on Artificial Intelligence, IBERAMIA 2004*, pages 147–154, Puebla, Mexico, November 2004.

Diana Santos, Paula Carvalho, Hugo Oliveira, and Cláudia Freitas. Second HAREM: new challenges and old wisdom. In António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira, and Paulo Quaresma, editors,

Computational Processing of Portuguese Language, 8th International Conference (PROPOR'2008), September 8-10, Aveiro, Portugal. Proceedings, number 5190 in LNCS, pages 212–215. Springer, 2008.

Diana Santos, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, Johannes Leveling, and Yvonne Skalban. GikiP at GeoCLEF 2008: Joining GIR and QA forces for querying Wikipedia. In Carol Peters, Tomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F.Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, and Viviane Petras, editors, *Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*. Springer, 2009.

Nuno Seco and Nuno Cardoso. Detecting User Sessions in the Tumba! Web Log. Technical Report. `http://eden.dei.uc.pt/~nseco/tumba.pdf`, March 2006.

Mário J. Silva, Bruno Martins, Marcirio Chaves, Ana Paula Afonso, and Nuno Cardoso. Adding Geographic Scopes to Web Resources. *CEUS - Computers Enviroment and Urban Systems*, 30(4):378–399, 2006.

Mário J. Silva. The Case for a Portuguese Web Search Engine. In *Proceedings of ICWI-03, the 2003 IADIS International Confere nce on WWW Internet*, pages 411–418, Algarve, Portugal, 2003.

Amit Singhal. Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4): 35–43, 2001.

Pedro Veiga and Diana Santos. Contributo para o processamento computacional do português: o CRdLP. In Maria Helena Mira Mateus, editor, *Mais Línguas, Mais Europa: celebrar a diversidade linguística e cultural da Europa*, pages 103–109, Lisboa, 2001. Colibri.

Jinxi Xu and W. Bruce Croft. Query Expansion Using Local and Global Document Analysis. In Hans-Peter Frei, Donna Harman, Peter Schäuble, and Ross Wilkinson, editors, *Proceedings of the 19th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'96)*, pages 4–11, Zurich, Switzerland, August 18-22 1996.

Appendix

A Survey on Geographic Information Retrieval

A Survey on Geographic Information Retrieval

Nuno Cardoso
Faculty of Sciences, University of Lisbon, LASIGE
ncardoso@xldb.di.fc.ul.pt

Abstract

This document surveys the geographic information retrieval (GIR) area, a specific information retrieval task concerned with document retrieval for geographically-related queries. Note that this is a preliminary version of the survey, with the purpose of supporting the PhD qualification proposal of the author; this survey will be further improved and updated, and it will be published when it reaches the quality requirements of computer science surveys.

This survey will introduce the main challenges of GIR, overview its main milestones and projects, present knowledge resources used in GIR research, and dissect its core steps, spanning areas such as named entity recognition, toponym resolution, geographic indexing and ranking, and query reformulation.

GIR research nowadays is being greatly fostered by the popularization of Wikipedia and other linked data sources such as DBpedia or Geonames.org, which provide a comprehensive and reliable amount of structured geographic information where GIR systems can base their geographic knowledge needs. We can denote an increase of GIR approaches that explore such services and resources for placename detection and disambiguation, toponym resolution or assigning geographic scopes to documents and queries, and for query reformulation approaches that are focused on understanding user queries and their search patterns.

In summary, GIR research looks well trailed to provide search tools that can effectively comprehend the user's needs within their geographic area of interest, narrowing the gap between "what the user wanted" and "what the user said", one of the main challenges in IR nowadays.

1 Introduction

Search tools are definitely taking part of our daily lives as a way to satisfy our information needs. As our information needs become more elaborate and specific, posing new challenges for the retrieval systems, new search tools are more aware of the context of search terms to provide special services in specific domains, such as search for services within a certain geographic area, namely in the whereabouts of the position of a handheld device.

As the user expectations are increasingly moving from "give me what I said" to "give me what I want" [Singhal, 2008], IR systems are closing in to the vision of the Semantic Web [Berners-Lee et al., 2001], researching for more reliable ranking approaches based on message understanding and in search contexts, rather than in frequencies of terms or document structure elements [Allan et al., 2003].

This report is a survey on geographic information retrieval (GIR), a subtask of information retrieval focused on retrieval approaches that provide a better search experience for user queries with a geographic area of interest. With the proliferation of information accessible to users, the geographic context is a common criterion of relevancy used by many users, that might find a document to be much more relevant if its geographic scope is within his area of interest. This is very common in searches for news, products, events or services that are located in the user's vicinities, and reflects the significative amount of geographically related queries that are submitted to web search engines (for instance, Kohler [2003] reports that one fifth of user queries have geographic terms).

1.1 GIR challenges

Although IR is the backbone of GIR, there are some challenges regarding storage and access of geographic information, and structured retrieval of geographic metadata. The extraction of such geographic metadata from documents also poses common challenges from the natural language processing domain. Overall, GIR addresses the following challenges:

- How to model the knowledge of the geographic domain, as conveyed by humans, in a way that can be used by software components of GIR systems?
- Which resources are best suited to be used to build such knowledge base, provided that the geographic knowledge must be accurate, actual and comprehensive?
- How to effectively capture the underlying subjects and geographic metadata from documents and queries? And how can this extracted information be used by a

ranking scheme that models the human notions of topic relevance and geographic similarity?

- How can the user interface of a GIR system potentiate the search experience and help the user in defining his/her particular information needs, and visualise the results?

1.2 Anatomy of GIR

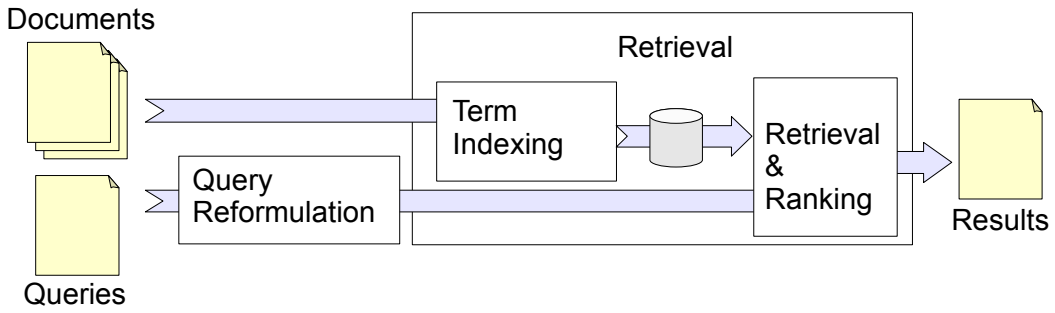

Figure 1: Generic architecture of an IR system.

Figure 1.2 illustrates the architecture of a classic IR system (for more information about IR, we recommend the books of Baeza-Yates and Ribeiro-Neto [1999] and van Rijsbergen [1979], and the papers of Arasu et al. [2001] and Singhal [2001]; the PhD thesis of Aires [2005] analyses the relation between what users want and how IR systems cope with that). The documents are stored and fed to a term indexing module, generating a term index that will be used by the retrieval and ranking module. The queries are sent to the retrieval and ranking module, which selects documents that match the query terms and ranks them according to a similarity metric that attempts to approximate the human's notion of relevancy. The queries may be preceded by a query reformulation step, whose purpose is to enhance the query and attenuate the terminological gap that exists between the vocabulary of the documents' authors and the users' queries.

While new GIR systems are being built and research groups are still trying different approaches, the overall GIR challenges remain the same; therefore, as a modern GIR system should tackle all of them efficiently, their generic architecture is presented in Figure 1.2

The main differences of GIR systems compared to classic IR systems are the following:

- IR systems typically use the existing document's text and structure to estimate relevance. GIR systems have to automatically generate additional geographic

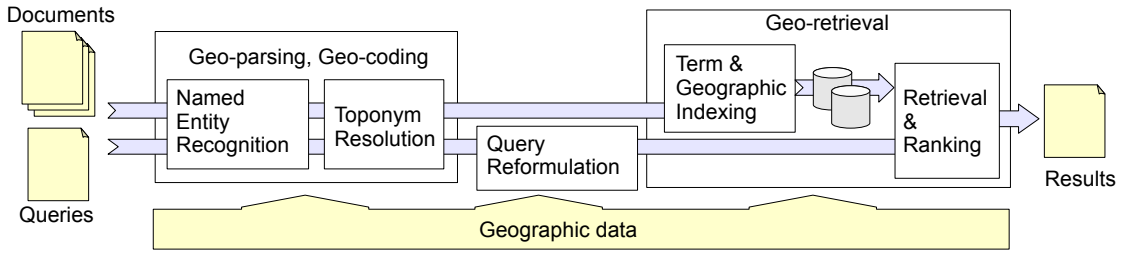

Figure 2: Generic architecture of a GIR system.

metadata from documents, thus having a more computational-demanding collection processing step.

- IR ranking modules uses term similarity measures to compute document relevance against a user query. In GIR, the ranking module computes relevance according to i) the similarity between subjects of documents and queries, and ii) the geographic similarity between document scopes and query scopes.

The main differences of GIR systems compared to structured retrieval systems are the following:

- GIR systems are concerned with the retrieval of unstructured documents, using relevance measures to estimate a relevance score of each document against a query, its subject and geographic scope. Structured retrieval systems are focused on the retrieval of structured data, and in the case of DB retrieval, the results are not ranked by relevance (they just have to satisfy the query criteria).

- The queries in GIR are issued in natural language, and thus are not queries aimed for relational data (as in DB retrieval) or structured information (as in most XML retrieval approaches) systems. The GIR system has the onus of capturing the information need and geographic area of interest for each query.

The main differences of GIR systems compared to geographic information systems (GIS) are the following:

- GIS aims to store and represent geographic data in an unequivocal way, often associating its entries with well-defined boundaries and geometric shapes. GIR systems have to cope with geographic criteria that are difficult to describe in an objective way (as in “north of Lisbon”), and use fuzzy notions of such geographic criteria to compute geographic relevance.

- The data of GIS is submitted to a strong curation and validation process to guarantee the precision of its geographic data. While GIR systems may explore these data, they still have to gather their own geographic data from the documents automatically (as such information is rarely available in an explicitly way), and it is unfeasible to manually validate such extracted geographic data.

1.3 GIR terminology

This section defines the GIR terminology that will be used in this report. The terminology adopts some of the definitions of other researchers and projects (cited when appropriate), and also includes the authors' point of view. Figure 3 schematises the terminology adopted for some of the entities, tasks and concepts in GIR.

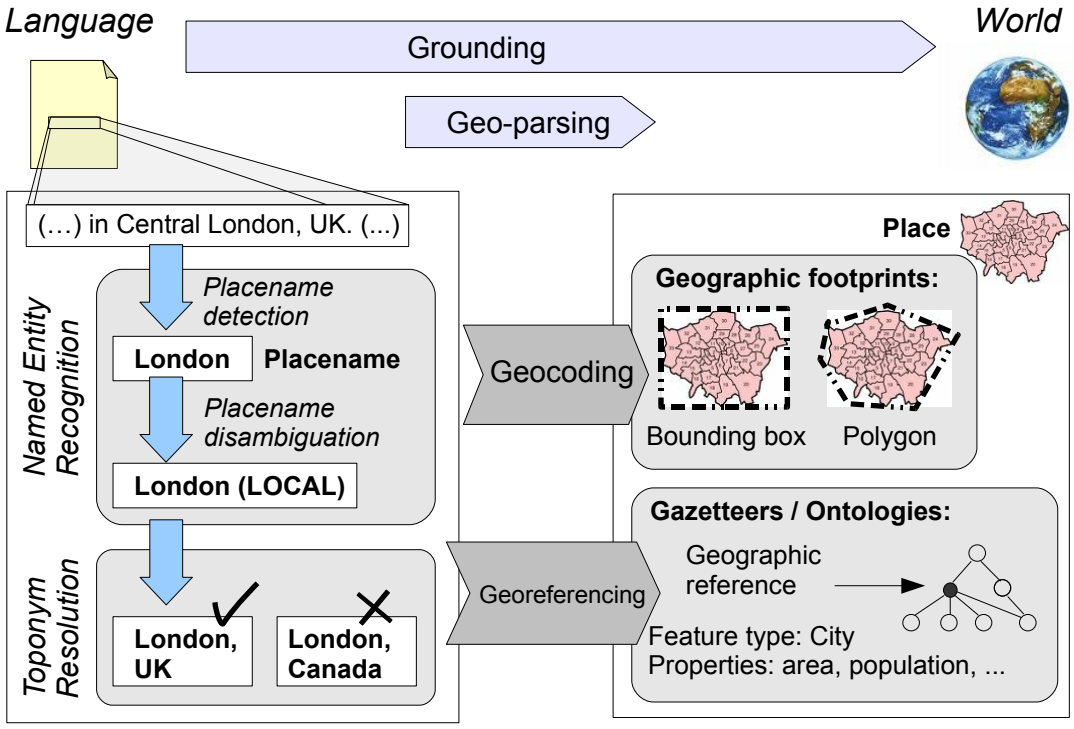

Figure 3: Schematization of GIR terminology.

1.3.1 Entities

Place - A *place* refers to a given spatial area from the Earth's surface, either motivated by its geo-political relevance, or by its physical characteristics (places are also referred in other works as *locations* or *geolocations*, and in an ontology conceptualization point-of-view, as a *geographical concept* [Jones et al., 2001]). See Bennett and Agarwal [2007] for a more thorough discussion about the meaning of "place".

Places typically contain important landmarks that justify their designation through one or more *placenames*. A placename (also designated as *toponym*, *geographic name*, or simply *geoname*) is a linguistic representation of a given place, and as such, its role depends on culture, language and contextual information [Santos and Chaves, 2006]. In summary, a place can have one or more placenames (as in "New York", "NY", "Big Apple"), and a placename may designate one or more places (as in "Springfield") and even concepts that are nor places (as in "Washington"). Still in the language domain, a geographic *type* represents a category of places that share similar characteristics, such as rivers, mountains, countries or cities. Geographic types are commonly used in the discourse to disambiguate placenames, as in "Washington state", although this does not completely solve the ambiguity issue.

Footprint - GIS concerns precisely on ways to represent places in an unambiguous way, avoiding the linguistic ambiguity. In most GIS systems, places are commonly associated to geometric shapes over a map of the Earth's surface. One great advantage of using geometric shapes is that they allow easy calculations for basic geographic operations such as distance, overlap or adjacency between places; on the other hand, they cannot capture precisely places whose boundaries are not crisply defined (as in "Midlands").

A *geographic footprint* of a place is the set of geometric shapes that define a given place [Hill, 2000]. Geographic footprints are normally represented by points or group of points; points may approximate the coordinates of the whole place, or they can represent the centroid of the geographic area of the place. Groups of points normally represent either open shapes (as in rivers), or closed shapes. *Polygons* are a typical choice to represent geographic footprints, as they can surround the boundaries of any area with a list of points and give a good approximation of the shape of any place. *Bounding boxes* are the most basic polygons, reduced to a rectangle representation of any place, requiring just two points to be defined. On the other hand, bounding boxes do not represent some places appropriately, introducing significant error on some geographic operations. For instance, Andrade and Silva [2006] showed that the shape of Spain generates a bounding box that includes the bounding box of Portugal, which may lead to erroneous conclusions that there is a significant overlapping area between these two places, when in fact there is none.

Geographic features - ISO 19109 [ISO 19109] introduces the notion of *geographic features* as geographic concepts that are uniquely represented in a resource such as a geographic ontology or a gazetteer. Geographic features can be associated to a given *geographic feature type* (as in country, island or continent), and characterised by an objective set of properties (or *geographic attributes*).

1.3.2 Concepts

Geographic scope - A *geographic scope* (or *geo-scope*), was first defined by Ding et al. [2000] as the geographic area that the creator of a given resource r intends to reach. Silva et al. [2006], on the other hand, defines geographic scope as a region, if it exists, whose most readers would find the document more relevant than the average reader.

From a GIR point of view, both queries and documents have geographic scopes: *query scope* refers to the geographic area (or areas) which the user is more interested to read about certain subjects, while the *document scope* refers to the geographic area (or areas) that are strongly related to the subjects addressed in the document. For instance, a document describing touristic attractions of the city of Lisbon is likely to be more relevant for an user that issued a query "tourism in Portugal" rather than an user which queried "tourism in Asia", because the document scope matches the query scope of the first user.

Geographic scopes can not be objectively determined, as they depend on the real intentions on the authors of documents and queries (even the user's acquaintance to the specified places is important for the definition of scopes Shanon [1979]), but can be approximated. Jones et al. [2004] refers to *document footprint* to designate the geographic footprint that best describes the geographic scope of a document. Cardoso et al. [2008], on the other hand, define *geographic signature* as a means to capture the geographic scope through a list of geographic concepts mapped by grounded placenames found in a document or a query, and weighted according to their importance for the geographic scope.

Geographic relevance - The geographic relevance of a document in respect to a given query represents the degree of concordance between document and query scopes. As for term relevance (see Figueiredo [1978] and Mizzaro [1997] for more information about relevance), the geographic relevance is a human notion of geographic pertinence between two geographic scopes.

1.3.3 Tasks

Named entity recognition (NER) - refers to the task of identification and classification of named entities (NEs) in the text. *Named entities* are names that refer to persons, places,

organizations, events or other relevant entities. NE identification (or detection) defines the terms that compose each NE, while NE classification assigns a semantic meaning to the NE.

Disambiguation - refers to the selection of the correct meaning for a given textual expression. Wacholder et al. [1997] mentions two types of ambiguity found when disambiguating MEs: i) *Structural ambiguity*, where the correct boundaries of NEs are hard to determine (for instance, Massachusetts Institute of Technology), and ii) *Semantic ambiguity*, where the referred subject is not clear (e.g. "Jordan" – is it a country or a person?).

Placename detection - refers to the identification of NEs in the text that *may* refer to geographic places.

Placename disambiguation - refers to the task of deciding whether a placename is a reference to a geographic place or not. The placename disambiguation task also addresses the *metonymic ambiguity*, where placenames are used metonymically to refer to another type of entity (for example, using "Brussels" to denote institutions of the European Union); while "Brussels" can be disambiguated to its place as the capital of Belgium, it does portray a different meaning, often a non-geographical one.

Toponym resolution - Leidner [2007, pp. 3] coins the term *Toponym Resolution* to designate the task of disambiguating the place referred by the placename, among all possible places (as in Cambridge in UK, or Cambridge in Massachusetts, USA), that takes place after placename disambiguation. In his own words, *"computing the mapping from occurrences of names for places as found in a text to a representation of the extensional semantics of the location referred to (its referent), such as a geographic latitude/longitude footprint"*.

Grounding - In a wider way, *grounding* refers to the process of connecting concepts presented in the text to its counterpart concepts in the real world. This cognitive process is very complex, as it requires a full knowledge of the world domain and on the language. In a way, the research on automatic 'grounding' techniques is one of the main goals of natural language processing systems.

From a GIR point of view, the grounding step is only concerned with the geographic facet of the documents and of the world, and thus assigning placenames to their geographic places. Leidner [2007, pp. 32] presents a diagram for the grounding task for GIR

(designated as *spatial grounding*), encompassing the steps of geo-coding, toponym resolution, and geographic expression resolution.

Geo-parsing - *Geo-parsing* can be understood as the step of performing spatial grounding on placenames in the text. The focus here is on having to parse text as input for the grounding process (mind that geographic evidence can be conveyed in other ways, such as server IP address or HTML `<META>` tags[1], which does not require text processing).

Geo-referencing and geo-coding - *Geo-referencing* is the core of geographic information systems (GIS), and consists on matching a given entity (as a physical object, a document, or a place) to a corresponding physical space that is unequivocally defined through a set of geographic properties. In GIS, geo-referencing is an elaborated and fine-grained task, spanning areas such as geodesy for its accurate measurement of geographic location, to characterise each geographic concept in the most accurate way. In GIR, geo-referencing denotes the assignment of unique geographic identifiers to the previously captured and disambiguated placenames.

Some authors define the task of *geo-coding* as the task of mapping implicit geographic data into explicit geographic representations [Leidner, 2007]. Under this definition, the toponym resolution task may be seen as a kind of geo-coding task. The difference between geo-coding and geo-referencing is that, while geo-coding is often related to the assignment of placenames to data other than text within a certain spatial model (as in polygons, images or maps), geo-referencing is used to designate the assignment of placenames to unique identifiers from a reference list (given, for instance, by gazetteers or ontologies). [Hill, 2006]'s book is a comprehensive reference on the geo-referencing subtask.

Geo-location - *Geo-location* is the task of inferring a given spatial location through the position of a device used to interact with a GIR system. Geo-location can be made, for instance, through IP address look-up on WHOIS servers like RIPE (`http://www.ripe.net/`) or APNIC (`http://www.apnic.net/`), or by GPS information passed by a mobile device. The geo-location can therefore be used to act as a geographic context for certain queries such as the search for services around the user ("restaurants near me").

Geo-tagging - *Geo-tagging* is the process of adding geographic metadata to documents. It is known through popular services like Flickr (`www.flickr.com`), where users can assign tags to pictures which may contain toponyms, or selecting the exact spot where the photo was taken in a map, this introducing the geographic coordinates as metadata in

[1] `http://dublincore.org/documents/1997/09/30/coverage-element/`

the image file. GPS-enabled camera devices do also perform automatic geo-tagging by introducing coordinate metadata when the picture is taken.

Document geocoding - Martins [2008] refers to *document geocoding* as the process of assigning geographic scopes to documents, by analysing the grounded geographic information from each document and deciding on an unique geographic reference (and/or footprint) that represents its geographic scopes. From a GIR point-of-view, document geocoding can be seen as a type of geo-tagging, where the geographic metadata being added aims to give a concrete delimitation of the geographic area of interest for each document. One example is the Web-a-Where system, devoted to the geo-tagging of web pages [Amitay et al., 2004].

Geo-similarity - *Geo-similarity* is the task of computing a measure that approximates geographic relevance between geographic scopes. Typical geo-similarity measure algorithms are aware of specific geographic restrictions given in user queries such as spatial relationships (for instance, "north of Paris" implies that the user is not interested in documents whose scope is inside the city of Paris) and the placenames given (for instance, "north of France" widens the desired geographic scope of documents to a broader area than "north of Paris").

1.4 Survey structure

This report is organised as follows: Section 2 provides an overview of the most important milestones on the recent history of GIR, referring its challenges, approaches and recent accomplishments. Section 3 overviews the resources that are being currently explored for GIR reasoning over the geographic domain. Section 4 overview the field of named entity recognition, a key step for placename disambiguation. Section 5 details the toponym resolution step. Section 6 presents current approaches for indexing geographic metadata and compute geo-similarity measures between documents and queries. Section 7 surveys the topic of query reformulation, from a GIR point-of-view. Section 8 concludes this report with a personal overview of the directions and challenges that GIR will face in the near future.

2 GIR systems

As any other inter-disciplinary research area, GIR started with a timid realization of the need for a geospatial-oriented retrieval of documents, with some initial drafts of the major GIR challenges and some prototypes [Larson, 1996]. IR grew up considerably after 2000, in line with the exponential growth of the world-wide web, and GIR soon started to have its own share of attention given the economic interest in search services that can provide personalised results according to the user's preferences.

The rise of GIR as a research area that spans information retrieval, data retrieval and geographic information management answered the need of restricting document searches to a given geographic area of interest, as users tend to find documents more relevant if they refer about products or services that are available and relevant around their whereabouts. This overview of the GIR history is therefore divided into three main stages:

The early days, when GIR researchers were having their first contacts with the difficulties of the task. While, in the theoretical plan, researchers were still figuring out the best way to model the geographic domain, in the practical plan they tend to adopt the simplest and easy way of extending a proven IR system with basic geographic data in the form of maps and coordinates to bootstrap their GIR systems, a trend that was understandable due to the limitations of computational capacity back then.

The emergence, where the first GIR initiatives started to crystalise into GIR proof-of-concepts for some experiments. The theoretic foundations of GIR were more evident now, and the time was ripe to establish GIR as an independent exercise from IR, needing its own evaluation initiatives, resources and strategies for its growth. The first commercial and academic GIR systems saw the daylight, and the appearance of (new) large geographic resources such as ontologies and gazetteers fostered GIR research.

GIR systems evaluation, where the birth of an annual GIR workshop and evaluation contest had a major impact on by bringing together a research community focused on GIR problems, fostering some recent research breakthroughs. The GIR research status today looks promising, with several papers and PhD thesis published around GIR, the exploration of new resources such as Wikipedia or geographic ontologies, and an active research community with means and tools to evaluate their progress.

2.1 The early days: the fist GIR proof-of-concepts

The PhD work of Hill [1990] focused on the use of geographic concepts for the retrieval of online bibliographic files within the earth sciences domain, aiming at evaluate the effectiveness of such geographical indexing approach. Hill concluded that the coordinate system was the best choice to compute geo-similarity and represent geographic scopes, as terms are too ambiguous and unpredictable (regarding the author's choice of words) to be used on geographic indexes.

Geographic coordinates were a natural choice to model the geographic domain and to bootstrap GIR research, as they are unambiguous representations of places, they can be easily manipulated by low-CPU powered computers to compute geographic areas and distances, they are easily matched to maps from user interfaces, and such data is easily made accessible due to the popularity of GIS systems.

While the conclusion that terms are not adequate to represent places leaves no objection, the fact is that people convey and write about geographic places in their native language [Egenhofer and Mark, 1995], and so it is unfeasible to ask for documents to have their placenames already geo-referenced to their geographic coordinates, and ready to be used by any GIR system. Add this to the increasing size of the collections, and the straight-forward conclusion is that there is a need to research strategies to automatically geo-reference placenames.

2.1.1 GIPSY

As far as we know, the GIPSY (**G**eoreferenced **I**nformation **P**rocessing **SY**stem) [Woodruff and Plaunt, 1994] is the first system to implement an automatic document geo-referencing process. Although it did not have a retrieval module, it is in our opinion the first GIR system that deserves its name. The GIPSY architecture used a basic thesaurus matching system, where text patterns, such as "University of California", were geo-coded to their corresponding real-world polygons, helped by the geographic gazetteer from the US Geological Survey's Geographic Names Information System (GNIS - `http://geonames.usgs.gov/domestic`).

The GIPSY system used polygons as geographic footprints of places in documents, and a simple overlay of polygons gave a weighted overview of the document scope, a process that was reported as CPU-intensive, even considering that the experiments with GIPSY were confined to the study area of California.

2.1.2 Defining GIR

Larson [1996] later coined the role of geographic information retrieval as a specialization

of IR concerned with "indexing, searching, retrieval and browsing of geo-referenced information sources, and the design of systems to accomplish these tasks effectively and efficiently." He cites the advantages of coordinates to encode geographic scopes of documents rather than placenames, and the need of investigating automatic ways to compute such coordinates. Larson's research envisaged a retrieval system over a geo-referenced digital library, so that the search process for geographically-related information could be much easier for end users.

Another important point raised by Larson is the notion that spatial relationships of geographic queries can be divided into geometric and topologic types, where the geometric type contains evidence for simple geographic calculation (as in "20 km north of"), and the topological type does not have a measurable distance or direction to base geographic calculations (typical in proximity queries such as "in the surroundings of").

2.2 The emergence: the WWW growth and the first GIR projects

2.2.1 GeoSearch

The end of the 90's saw the consolidation of the world-wide web as an amazing information source, with an uncontrollable growth rate. In this context, Buyukkokten et al. [1999] seized the opportunity to exploit the geographical location information of websites to improve web search engines, as an additional relevance criteria for document ranking. Their work focused mostly on the detection of overall geographic scopes for some websites, through capturing placenames and zip codes in their pages, and mapping host IP addresses into their corresponding geographic coordinates (this heuristic is applied with caution, as most host IP addresses may not be strongly related to the website's geographic scope).

In their work, Buyukkokten et al. also reported that some sites can have a global scope of interest, although they have a defined place of origin (for instance, the pages from the website of the "New York Times" are not relevant only to persons from the New York City area). This subject is later detailed by Ding et al. [2000] and Santos and Chaves [2006], which states that there is a significant difference between the placenames found on a document, and the scope of interest (to the readers) for that same document, and GIR systems should be aware of this offset.

The work of Buyukkokten et al. [1999] and Ding et al. [2000] culminated on the GeoSearch search engine (`http://geosearch.cs.columbia.edu`), a proof-of-concept system that estimates geographical scopes for newspaper sites using the incoming web links distribution. While it is not a full GIR system and key issues such as geographic indexing & ranking are still unaddressed, their research has the merit of exploring a basic

US topological hierarchy and the web link structure to determine geographic scopes of documents (or websites), avoiding the use of geographic coordinates.

McCurley [2001] picked up the work of Buyukkokten et al. [1999] and focused his research on the development of a navigation tool for browsing webpages by geographic proximity rather than other similarity measures. His approach also uses the geo-parsing of documents by looking up addresses, phone numbers, zip codes and placenames given by the GNIS gazetteer and the Geographic Names System from the United States National Imagery and Mapping Agency's (NIMA)[2]. He also explores link structure to geo-reference documents.

McCurley's GIR prototype features a browsable map that presents URL lists according to geographic proximity, thus addressing the problems of building intuitive interfaces to display geographically-ranked content. The author concluded that exploring geographic content from web documents is a feasible and promising task.

2.2.2 GeoVSM

The aforementioned proof-of-concept prototypes often neglected key GIR issues such as geographic retrieval and ranking, where a common geospatial model must be established in order to compare and evaluate documents and queries according to their geographic proximity. Unless the user is (unlikely) willing to give the coordinates or draw the polygon of the area that he is interested in, GIR systems cannot avoid the fact that they must infer geographic scopes from the terms of queries and documents.

The work of Cai [2002] on geo-libraries, even thought primarily concerned with merging map and text approaches, has the merit of addressing the problem of merging geographic proximity and term similarity into a common ground for an effective computation of relevance for document retrieval (thus justifying GIR as a midway between GIS and IR). Cai proposed the GeoVSM model, a geographically-inspired version of the well-known vector space model (VSM) used on IR systems [Salton et al., 1975]. The GeoVSM models the search space into two subspace models: a i) thematic subspace, that models the different topics addressed in an n-dimensional space of terms of the VSM, and ii) a geographic subspace, that models the spatial domain as a geographic coordinate system, where document similarity is given by the proximity of the coordinates. The GeoVSM approach generates two ranking scores (one for each of the subspaces), which are afterwards combined through a linear equation that generates a single ranking score. Yet, Cai's work does not explain how the geographic coordinates of documents and queries can be extracted, in order to be employed in the geographic subspace.

[2]Now part of the U.S. National Geospatial-Intelligence Agency.

2.2.3 MetaCarta

The geographic search engine MetaCarta (`www.metacarta.com`) is a commercial GIR system with a strong focus on placename disambiguation. MetaCarta uses machine learning approaches on the context surrounding placenames (postal codes, for example), specific term patterns around placenames, and simple heuristics, such as ranking place importance through population count, to estimate a confidence level measure during placename disambiguation [Rauch et al., 2003].

As a commercial search engine, MetaCarta is also concerned with geo-parsing large collections of documents and with retrieval efficiency and effectiveness with respect to geographic and subject criteria, but details on such approaches are not available. Document relevance is given through a balanced combination of a modified term weighting score (not specified, but referred as based on a standard technique cited in Robertson and Spärck Jones [1997]) and a *georelevance* score that uses the calculated confidence levels for grounded placenames in documents.

MetaCarta provides a GeoTagger Web Service (`http://developers.metacarta.com/api/geotagger/`), where users can submit documents and it geoparses all placenames and returns tagged information about the grounded placenames, along with a degree of confidence for each placename geocoding.

2.2.4 Web-a-Where

Web-a-where is a geo-tagging system developed in IBM by Amitay et al. [2004]. The goal of Web-a-Where is to assign *geographic focus* (that is, geographic scopes) to web pages, using the WebFountain data mining framework system [Gruhl et al., 2004]. As a work concerned with processing large collections of documents, Web-a-Where preferred simple and fast heuristics for placename disambiguation, rather than slower NLP-based algorithms.

Web-a-where features a gazetteer that contains an hierarchical representation of the most important places in the world (countries plus cities with more than 5,000 inhabitants). Placename disambiguation is based on the taxonomical location of the placename in the gazetteer, which approximates the human notion of importance of places according to its type, and that it is more likely that a given placename refers to the most important geographic place.

Web-a-where's geographic scope assignment algorithm uses the grounded placenames and the gazetteer hierarchy to decide a final list of regions that best describe the scope of each document. The algorithm takes in consideration the importance of the various taxonomy levels of the gazetteer, and the hierarchical relationships between the grounded places to come up with the final geographic scope. Their evaluation with a

corpus of 20,000 webpages was taken from the ODP: Regional (`http://dmoz.org/regional`) reports a 92% accuracy on scope assignment to the country level.

2.2.5 SPIRIT

The SPIRIT project (`www.geo-spirit.org`) is an EU-funded research project devoted to the design and development of a working spatially-aware web search engine [Jones et al., 2002]. With this ambitious goal, SPIRIT had to address all GIR challenges and compensate the lack of specific geographic resources, namely: i) designing and creating a geographic ontology suitable to provide a geographic knowledge base for all GIR components, ii) indexing geographic metadata gathered from automatic geoparsing of web documents, iii) capturing the geographic criteria and spatial relationships from user queries, and iv) developing a suitable interface so that users can browse spatially-relevant results and redefine their geographic restrictions, for instance by interacting with a digital map.

More details on the implementation of the SPIRIT search engine can be found in Jones et al. [2004]. The whole search engine revolves around the geographic ontology, as it is used to detect geographic criteria from queries (treated as $<$ *terms*, *spatial relationship*, *place* $>$ triplets), perform basic query term expansion, disambiguate placenames, geo-parse documents and map them to geographic footprints (centroids, polylines and polygons), for both queries and documents.

The relevance ranking is given by combining the score of BM25 text weighting scheme [Robertson et al., 1992] with the spatial distance between the footprints of document scopes and query scope. Once again, the ontology is used to compute scores for specific spatial relationship, such as orientation ("north of"), adjacency or overlapping. No further details were given on how SPIRIT combines the scores.

2.2.6 GReaSE

The GReaSE Project (`http://xldb.di.fc.ul.pt/wiki/Grease`) started in 2004 with the purpose of researching methods to provide geographic reasoning on geographic web search engines [Silva et al., 2006]. GReaSE's methodology involved the development of a comprehensive geographic knowledge base from public resources [Chaves et al., 2005], allowing the generation of a fine-grained geographic ontology that will assist the components of the search engine. GReaSE pays special attention to the assignment of geographic scopes to documents, which is performed through a graph-like algorithm that used grounded placenames to select a single encompassing geographic reference [Martins and Silva, 2005].

GReaSE's approach is also notable for its focus on georeferencing documents to geographic concepts in an ontology, rather than geocoding to geographic footprints [Martins et al., 2007]. This approach means that the proposed indexing and ranking modules do not use a spatial model based on coordinates, but rather rely on reasoning through the ontological relations between geographic concepts. While this approach requires more CPU-expensive algorithms for geoparsing documents, it is more robust for geographic queries with fuzzy placenames that are difficult to represent through geometric shapes.

The follow-up project, GReaSE II, started in 2008, and focuses on the automatic generation of geographic signatures for documents, the automatic query reformulation of geographic queries, the fusion of geographic and textual rankings, a better support for multi-lingual documents, the addition of physical geography of the world in the knowledge base, and the development of faceted interfaces for GIR.

2.2.7 Google local and Yahoo! local

Also worth mentioning is the creation of location-aware services by Google (`local.google.com`) and Yahoo! (`local.yahoo.com`), which work as a yellow-page service where users can search for previously georeferenced services. While these web services are not true GIR systems, in the sense that they do not geoparse the documents and thus do not have the capability of returning geographically-located documents regarding any given subject, they nonetheless address some common challenges for GIR, namely the development of a suitable interface to present georeferenced results, and handling and disambiguating geographic queries to their grounded places.

2.3 GIR systems evaluation: the growth of a GIR community

Many research areas related to computer science experienced a significant boost on their achievements by gathering the research community around specific tasks and common goals. For instance, the Message Understanding Conferences (MUC) had a significant impact on fostering the field of information extraction, motivating the community around periodical evaluations on challenging objective tasks of automatic message understanding [Hirschman, 1998].

Likewise, GIR blossomed with the organization of periodic workshops and evaluation contests, allowing GIR researchers to present and evaluate their approaches, providing valuable feedback to perfect their work. Also, some prototype geographic search engines emerged, namely for the German [Markowetz et al., 2005] and the Portuguese communities (with GeoTumba, `local.tumba.pt`).

2.3.1 Workshops

The fist GIR workshop was the workshop for Analysis of Geographic References, held along with HLT-NAACL in May 2003 in Edmonton, Canada [Kornai and Sundheim, 2003]. The works presented focused on placename detection, the use of context and place types for semantic disambiguation of placenames, and toponym resolution.

Purves and Jones [2006] organised a workshop on Geographic Information Retrieval in 2004, held along with the SIGIR conference in Sheffield, United Kingdom. The idea was fueled by the organisers' involvement on the SPIRIT project, and it is still being organised every year.

A workshop on Methodologies and Resources for Processing Spatial Language was held along with LREC in 2008 [Katz et al., 2008]. The goal of this workshop was to promote works on the standardization of resources for processing spatial language. Following the ideas raised by Egenhofer [2002], envisaging a geospatial web where a normalised representation of geographic entities would greatly benefit GIR research, the workshop focused on "*methodologies for mapping natural language expressions that describe locations, orientations and paths to the geospatial entities they refer to and for encoding the spatial relationships among the entities described*". Among current markup languages to encompass geographic informations are the SpatialML (`http://sourceforge.net/projects/spatialml`), the Open Geospatial Consortium's GML (`http://opengis.net/gml/`), KML (`http://www.opengeospatial.org/standards/kml/`) and TRML, the proposal of Leidner [2006] for a toponym resolution markup language.

The workshop on Location and the Web (LocWeb2008), held at the WWW 2008 Conference in Beijing, China, addresses the problem of the lack of explicit spatial content on the Web information, by promoting works from the academia and industry focused on extracting and representing geographic data from web documents Boll et al. [2008].

The workshop on Geographic Information on the Internet (GII 2009 - `http://www.moromete.net/GII/`), held at the ECIR 2009 in Toulouse, France, on the other hand, is concerned on how to best explore the increasing amounts of geographic data from documents (for instance, user-generated data, such as tags), and use it as a knowledge resource and how to apply it for the retrieval and display of personalised search results.

2.3.2 GIR evaluations

System evaluation plays an important role in measuring the suitability of different GIR approaches, by objectively assessing and comparing their performance on a common task, and thus selecting the most fit strategy for the problem.

Woodruff and Plaunt [1994] first referred the need for an evaluation benchmark, namely with a comparison of GIR results against a set of manually-geocoded documents. Nonetheless, a fair evaluation base that suits most GIR systems and follows the Cranfield paradigm [Cleverdon, 1967] requires a considerable amount of human work to assess a representative collection of documents against geographically challenging topics.

Since 2005, the CLEF evaluation conference on cross-language IR [Peters and Braschler, 2001] organised a specific evaluation track for GIR systems, dubbed GeoCLEF. Throughout the series of four editions of GeoCLEF [Gey et al., 2006, 2007b, Mandl et al., 2008b,a], it is noticeable the improvements made to the evaluation task itself, namely the identification of topics initially presented by Santos and Chaves [2006] and that presented considerable challenges from a GIR point of view [Gey et al., 2007a], and the progress achieved by regular participants with several interesting GIR approaches.

Nonetheless, GeoCLEF had its own limitations and shortcomings; for instance, the document collection, based on news reports, prevents GIR systems to explore document metadata, available in many document collections. The geographic coverage of topics was in some editions under-challenging, which may explain the fact that GIR systems had many difficulties on observing improvements on retrieval performance when using geographic reasoning, being often outperformed by classic IR approaches [Kornai, 2006].

2.3.3 Related evaluations

A generic GIR evaluation task, such as GeoCLEF, can only give overall performance measures of a whole GIR system; as a complex system composed of several components that perform different tasks, one also needs specific evaluation benchmarks for each component in order to detect performance bottlenecks, identify failure points and redirect research efforts to alternative problem decompositions.

The named-entity recognition is actually well-suited with evaluation initiatives, namely the MUC conferences [Hirschman, 1998], the CoNLL shared tasks of 2002 and 2003 [Sang, 2002, Sang and Meulder, 2003] and the HAREM evaluation [Santos and Cardoso, 2007, Santos et al., 2008]. Regarding toponym resolution (and in some aspects, placename grounding), [Leidner, 2007] describes in his PhD thesis his own evaluation methodology proposal, using a subset of documents taken from former MUC and CoNLL collections, and specifically annotated for the task with TRML tags. While this is not yet an evaluation benchmark widely used by the GIR community, it is a worthy initiative to overcome the lack of any evaluation methodology for such task.

Query parsing also got a special attention on the 2007 edition of GeoCLEF, with a specific subtask [Li et al., 2007]. The task guidelines focused both on the identification of $<$ *what*, *spatial relationship*, *where* $>$ triplets and in the classification of query types (Map,

Yellow Page and Information) on a real search engine query logs with 800.000 queries. A follow-up track called Log Analysis and Geographic Query Identification (LAGI – www.uni-hildesheim.de/logclef/), still on the forge, will take place on CLEF 2009.

In the last edition of GeoCLEF, a pilot task concerning the retrieval of geographically-oriented answers from Wikipedia was also organised, entitled GikiP [Santos et al., 2009]. The main goal of GikiCLEF was to provide a sandbox for other geographic challenging topics more typically handled in question answering, and to use Wikipedia as a base collection, thus promoting more NLP-based strategies and strong geographic-reasoning for each retrieval. The follow-up task is now called GikiCLEF (http://www.linguateca.pt/GikiCLEF), being nowadays the only evaluation task that is concerned with GIR goals.

3 Resources for GIR

As GIR systems distinguish themselves from IR by their ability to understand and reason over geographic criteria, it is crucial to develop geographic knowledge bases that model the human notion of the geographic domain in a comprehensive and machine-friendly format. Human knowledge can be modeled in several ways, from loose and short tag clouds typically associated to folksonomies, through simple taxonomies written in natural language and structured in markup languages, to strict schemas that encode human knowledge in a rigid hierarchy of concepts and relationships. Figure 4 presents a diagram inspired on the presentation of Silva [2007], such ways vary in terms of machine-readability and objectiveness.

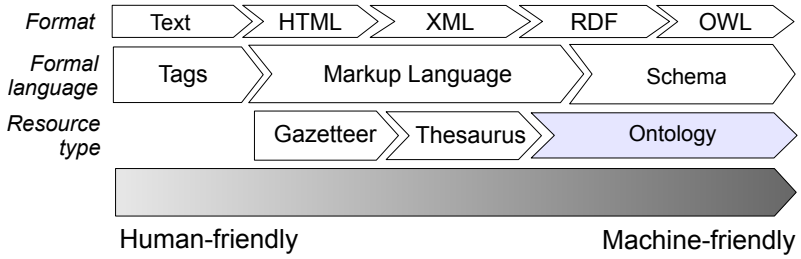

Figure 4: Models of knowledge representation

Ontologies - An *ontology* is a formal description of concepts and their relationships among them, and thus is the desiderata of the representation of human knowledge for computers. As portrayed by Berners-Lee et al. [2001] in his vision of the Semantic Web (SW), ontologies are built over layers that are formalizing knowledge and thus encapsulating it on a format that can be understood by machines: XML became the standard markup language, RDF is format used to describe resources and exchange knowledge between systems [Lassila and Swick, 1998], and above the XML/RDF layer, Smith et al. [2004] specify the Web Ontology Language (OWL), a knowledge representation language suited to describe classes and relationships. All these formats are W3C recommendation standards for knowledge representation.

The SW is a decentralised model, where knowledge can be partitioned among several domains and data can be scattered on the Web. The Linked data model ensures that the SW data is represented through dereferenceable URIs that can be interlinked [Berners-Lee, 2007]. Domain ontologies are ontologies dedicated to model a specific part of the human knowledge; *geographic ontologies* (also called geo-ontologies by other

authors, or as geospatial ontologies by the W3C[3]) are therefore domain-specific ontologies focused on describing the geographic domain. Building a geographic ontologies poses many challenge, namely the conceptualization of what humans understand as the major geographic categories and features [Smith and Mark, 2001].

Geographical ontologies are a fit resource for GIR, as they keep the quality and quantity of information that bases crucial GIR steps such as calculating geo-similarity scores for document ranking [Jones et al., 2001]. Throughout the rest of this document, we will shorten the references to geographic ontologies by referring to them as simply *ontologies*.

Gazetteers - A *gazetteer* designates a geospatial dictionary of geographic names [Hill, 2000]. In a figurative sense, a gazetteer performs the same function for the geographic domain as a standard dictionary does for the language domain, that is, provides a basic (but not complete) explanation of the spatial location of placenames.

Hill [2000] states that a gazetteer entry should have at least the following three constituents: i) the placename, with optional associated variants, ii) the geographic feature type, such as city, lake or country, and iii) the geographic footprint. The creation of gazetteers (and ontologies) follow a thorough process of curation and validation of raw information, to ensure a high degree of credibility on its geographic information (see Hill et al. [1999] and Axelrod [2003] for further details on building a gazetteer). As the boundaries, properties and even designations of places are constantly changing, it is also important that such resources are regularly updated.

Thesauri - A *thesaurus* can be understood as a dictionary that includes semantic relations between entries, such as synonyms or antonyms. From a geographic point of view, a thesaurus may be seen as an enhanced gazetteer, with basic relationship information between entries.

Ontologies, gazetteers and thesauri can be seen as three different representation types of the geographic domain, with distinct purposes; thesauri is a resource for linguistics, far from having the ontology's strict format as they are still a resource for human consumption. Gazetteers are a resource of GIS, concerned on the correct description of properties of places, and lacking the information about the relationships between entries and the hierarchical relationships between concepts. For instance, a gazetteer can readily provide a population count for the city of Brussels or the area of Belgium, but it does not provide

[3] `http://www.w3.org/2005/Incubator/geo/XGR-geo-ont-20071023/#ontologies`

a way to infer the relationship between these two geographic concepts. Finally, ontologies are a resource more suited for Artificial Intelligence research fields, providing the basic structures for reasoning over concept properties and relationships.

This chapter reviews the most popular ontologies, thesauri and gazetteers in the GIR community, as well as other raw resources from where knowledge can be mined and used, even though they may lack the structure and objectivity of the former kinds of knowledge resources.

3.1 Gazetteers

The available gazetteers differ mostly on their world coverage, their detail (or the number of entries) and on the geographic properties associated to each entry. While an useful gazetteer should have a decent coverage of the most relevant world landmarks, a very large gazetteer might include an excessive number of place candidates for each placename, and thus require robust toponym resolution approaches to handle such (noisy) data.

The first gazetteers reported to be used by the first GIR experiments were the U.S. Geological Survey's (USGS) Geographic Names Information System (**GNIS**) (`http://gnis.usgs.gov/`) and the National Imagery and Mapping Agency's Mapping Agency's (**NIMA**) Geonames server (now part of the U.S. National Geospatial-Intelligence Agency (NGA), and known as the GEOnet Names Server (**GNS**) `http://geonames.nga.mil/`). The GNIS gazetteer spans the US territory and includes 2 million entries, while the GNS gazetteer covers the rest of the world, and reports around 6 million entries.

The Alexandria Digital Library (ADL, `http://www.alexandria.ucsb.edu/gazetteer`) has nearly 6 million geographic locations around the world, and includes the datasets from the NIMA and GNIS gazetteers, plus other additional information such as the bounding boxes for administrative areas [Hill, 2000].

GeoXwalk (`http://www.geoxwalk.ac.uk/`) is an UK-focused gazetteer which includes geographic footprints represented by lines and polygons, aimed to provide a gazetteer service acting as a remote geocoding system and solving location-based queries such as "What parishes fall within the Lake District National Park" [Reid, 2003]. With a database schema based on ADL, it has over 530,000 entries[4] and even provides a geocoding service (not avaliable online) [Densham and Reid, 2003].

Geonames (`http://www.geonames.org`) is nowadays the most comprehensive gazetteer available, with over 6.5 million entries, and is gaining popularity due to its extensive list of web services (`http://www.geonames.org/export/ws-overview.html`) that perform basic operations such as "finding the nearest street"

[4] `http://edina.ac.uk/projects/geoxwalk/features.html`, values from March 2006.

or "finding the nearest Wikipedia page" for a given geographic coordinates. Geonames includes data gathered from the U.S. National Mapping Agencies, Statistical Offices, Postal codes and the National Geospatial-Intelligence Agency. Geographic data include placenames in several languages and population counts.

The **World Gazetteer** (`http://www.world-gazetteer.com`) is a gazetteer more concerned with population statistical data. Nevertheless, it provides data about geographic areas, coordinates, alternative placenames and population sizes to entries from several administrative division levels (from country to cities, towns and metropolitan areas).

3.2 Thesauri

The Getty Thesaurus of Geographic Names, (**TGN** - `http://www.getty.edu/research/conducting_research/vocabularies/tgn/`) [Harpring, 1997] is one of the most widely used geographic resources in GIR, with around 1,106,000 placenames of about 912,000 places according to TGN's Wikipedia page[5]. TGN is often described as a gazetteer, although Getty claims that TGN is a thesaurus since it contains hierarchical, equivalence, and associative relationships.

WordNet (`http://wordnet.princeton.edu`) is a lexicon database that has also features from thesaurus, describing concepts in natural language and grouping them into sets of cognitive synonyms, called *synsets* [Miller et al., 1990, Fellbaum, 1998]. The WordNet version 2.1 has over 80,000 synsets for around 117,000 unique nouns.

WordNet is exhaustively used by natural language processing applications for several tasks, spanning part-of-speech tagger, word sense disambiguation, text mining, among others. Originally in English, WordNet is being adapted to other languages, with initiatives as the EuroWordNet project [Vossen, 1998] and the Global WordNet Association (`http://www.globalwordnet.org/`).

The work of Buscaldi et al. [2006] explored WordNet for their GIR prototype system, mostly for the expansion of geographic terms on the queries and documents. While their first experiments failed to show significant improvements on GIR retrieval [Buscaldi and Rosso, 2008a], their work enhanced WordNet by automatic georeferencing some of the geographically relevant synsets with geographical coordinates fetched from the Wikipedia-World project, producing the **Geo-WordNet** [Buscaldi and Rosso, 2008b]. A prototype system called GeoWorSE (an acronym for Geographical Wordnet Search Engine) used the Geo-WordNet and reported an improvement of GIR performance, specially when using maps to filter the query scopes [Buscaldi and Rosso, 2007].

[5] `http://en.wikipedia.org/wiki/Getty_Thesaurus_of_Geographic_Names`, accessed January 2009

3.3 Ontologies

Following the W3C recommendations, the aforementioned Geonames also has released an ontology built on top of the existing gazetteer (`www.geonames.org/ontology/`), thus constituting probably the largest geographic ontology available nowadays. The Geonames ontology is encoded in OWL/RDF format, and it reports more them 6.5 million features and around 94 million RDF triples.

The **Geo-Net PT** is a geographic ontology initially crafted to cover the Portuguese territory, with detailed information from major administrative divisions to street level [Chaves et al., 2005]. The Geo-Net PT 01 version included data from online information sources such as the Portuguese postal codes service. It contains about 418.000 geographic concepts and with mostly part-of and adjacency relationships between them, the Geo-Net-PT 01 version spanned only the administrative domain. A Geo-Net PT version covering the most relevant places of the world (using data from Wikipedia and the World Gazetteer) was used to provide geographic knowledge for a GIR prototype that participated on GeoCLEF [Martins et al., 2007].

The Geo-Net PT 01 was evaluated and measured by Chaves and Santos [2006] against a web collection, concluding that web documents are rich in geographic placenames and are a valuable resource for enriching the contents of Geo-Net PT. Additionally, the majority of placenames in the Geo-Net PT are not used by the authors of the web documents. The PhD work of Chaves [2008] proposed precisely a method of ontology enrichment with geographic information extracted from the web, and the next version (Geo-Net PT 02) will include the physical domain, as well as more detailed data and a fine-grained feature type hierarchy [Chaves et al., 2007, Rodrigues, 2008].

3.4 Wikipedia

Wikipedia (`http://wikipedia.org`) is a free online encyclopedia launched in January 2001 by Jimmy Wales and Larry Sanger. Written by voluntary contributors around the world, Wikipedia has evolved to a massive Internet phenomenon, containing millions of articles in several languages with an impressive degree of accuracy, reliability and trustiness of its information.

In a way, Wikipedia can be seen as having characteristics from ontologies, gazetteers and thesauri. In fact, Wikipedia articles describing geographic places are unique and well disambiguated, with a short and concise description of the place in the first paragraph, and with infoboxes containing associated properties, such as feature type, area, population or geographic coordinates (actually, some georeferenced Wikipedia documents are automatically geotagged by Wikipedia robots that use the Geonames.org gazetteer

service as a data source)[6]. Although not as explicitly as an ontology, Wikipedia's link structure and categories provide a source of relationships among geographic concepts, and thus providing an excellent ground for researchers on information extraction areas. Overell [2009] is precisely focused on extracting geographic knowledge from Wikipedia and apply it for the benefit of GIR.

Although Wikipedia documents are well structured in HTML, following a Manual of Style[7] that has recommendations about using HTML elements such as headings and lists to organise information, it lacks the strict grammar and resource description strength given by RDF format. Wikipedia still requires a "semantic layer" over its contents, so that it can be a suitable machine-friendly resource.

Völkel et al. [2006] envisage a Semantic Wikipedia, where the Wikipedia documents can be enriched with semantic annotations, to facilitate information access for machines. While this initiative requires that such semantic annotations are to be introduced by the community, other approaches try to generate such information automatically. Wu and Weld [2007] propose an autonomous and self-supervising machine-learning approach to bootstrap Wikipedia and achieve such semantical desiderata. Another initiative worth mentioning is the PlaceOpedia (`www.placeopedia.com`), which provides a map interface so that users can connect Wikipedia pages to their locations,

DBpedia (`http://dbpedia.org`) is the best known effort on extracting information from the Wikipedia and representing it through RDF triples [Auer et al., 2007]. DBpedia's approach is to explore the infobox templates, which include properties and descriptors for the subject [Auer and Lehmann, 2007]. In November 2008, DBpedia reported that the dataset version 3.2 has around 274 million RDF triples generated from documents written in 14 languages, describing over 2.6 million entities, including 213,000 persons, 328,000 places, 57,000 music albums, 36,000 films and 20,000 companies.

DBpedia's datasets are freely available, and their information can be accessed through SPARQL queries. DBpedia is interlinked with several datasets, and according to the Linking Open Data project (`http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`), it plays a central authority role on knowledge exchange (see Figure 5).

YAGO (`http://www.mpi-inf.mpg.de/~suchanek/yago/`) is a similar initiative to DBpedia, with over 2 million entities and 20 million RDF triples [Suchanek et al., 2007]. YAGO's developers see it as an unification of Wikipedia and WordNet, resulting in a large ontology with great coverage and an accuracy rate reaching 95% on fact correctness. Unlike DBpedia, YAGO relies on Wikipedia categories to base its extraction

[6] `http://de.wikipedia.org/wiki/Wikipedia:WikiProjekt_Georeferenzierung/Wikipedia-World/en`, accessed on February 2009.

[7] `http://en.wikipedia.org/wiki/Wikipedia:ManualofStyle`

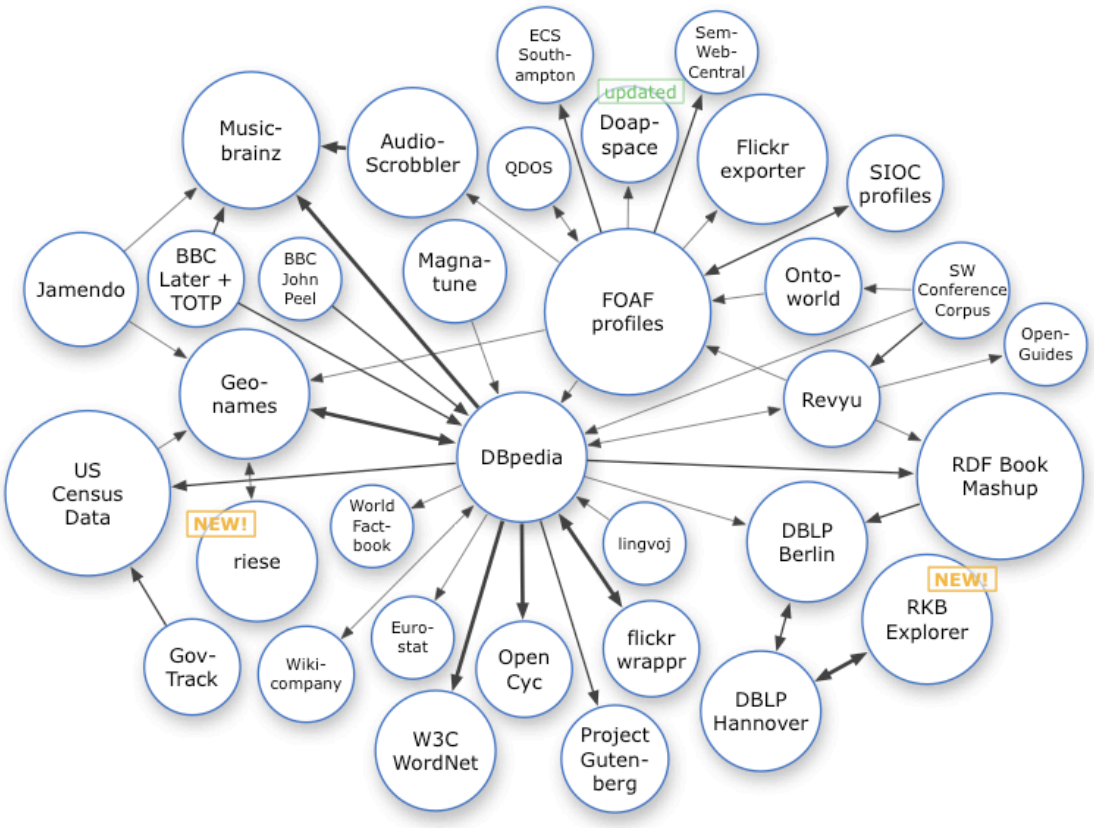

Figure 5: DBpedia featured on the center of the Linking open data dataset cloud, accessed at `http://richard.cyganiak.de/2007/10/lod/` in January 2009.

of facts, and it focus exclusively on English.

3.5 World-wide web

The world-wide web is arguably the largest information repository ever built, virtually encompassing all world knowledge through an impressive number of documents (texts, images and videos), which can be easily counted to billions. From an information extraction (IE) point of view, the advantages of the world-wide web are also its biggest disadvantages, since its chaotic and unstructured nature makes it difficult to extract information in an automatic way (for instance, Agichtein and Gravano [2000] report that their Snowball relation extraction system from plain-text collections would require several modifications to properly process HTML document). For more information on IE, we

recommend the survey of Cowie and Lehnert [1996].

The KNOWITALL initiative aimed to explore the world-wide web as a knowledge resource, relying on co-occurrence statistics and in the simple sense that the plausibility of a certain information may be supported by its scatterness across multiple distinct documents [Etzioni et al., 2005]. With simple text patterns as in "X, such as Y, Z" and exploring HTML markup such as lists, KNOWITALL can extract facts associated with a certain probability, which depends on the amount of evidence that supports each fact.

Recently, there has been some works around new IE paradigms on unsupervised strategies and the development of IE systems which require little or no tagged corpora or manual patterns , with shallower and faster approaches that make them more suitable to process very large corpora; Sekine [2006] proposed an "On-demand Information Extraction" paradigm, envisaging an IE system that automatically create patterns to extract relationships and are more adaptable to any type of topics of extraction. The work of Shinyama and Sekine [2006] refers to Preemptive Information Extraction as the automatic creation of information tables from all topics in advance without human intervention. Finally, Banko et al. [2007] introduce a new paradigm Open Information Extraction (OIE), where the IE system is expected to make a single pass on the corpus, and immediately extract a large set of relations without any human intervention. The TEXTRUNNER system follows the OIE paradigm, and it is able to process large quantities of Web text faster than KNOWITALL, with a better precision rate for the same recall values.

4 Named entity recognition

Named-Entity Recognition (NER) is a subtask of IE, with the purpose of identifying and semantically classifying proper names in the text. The main challenges of NER are related to the intrinsic vagueness / ambiguity of natural language, namely:

- the identification of named entities (NEs), that is, the correct definition of the boundaries of an entity. For instance, "São Tomé and Prínicipe" designates a single country while "San Marino and Italy" designates two distinct countries, for an identical entity structure.

- the semantic classification of NEs, that is, a basic categorization of the true meaning of the NE, according to a certain class hierarchy. For instance, "Washington" may designate persons, events, organizations or places, depending on its context. NE classification often comprises a large NE disambiguation step, to decide on the right meaning among all possible candidates.

In GIR, named entity recognition is mostly focused on the recognition of placenames and other types of entities that can be geocoded, such as postal codes. Nonetheless, such NER approaches still have to deal with the fact that most placenames may designate entities that have no geographic connection, and in a more broader picture, user queries may also refer to certain entities that can be therefore recognised and used on the retrieval process. For instance, in a query "Bill Clinton visits to Germany", by recognizing the person "Bill Clinton" a GIR system can have a better understanding on the subject, thus focusing query expansion based on that information and influence document ranking to promote documents referring specifically to such person.

4.1 NER approaches

The work of Wacholder et al. [1997] reports that NER disambiguation is mostly based on context and world knowledge, and predicts already how NER accuracy can be gained through the use of NLP techniques, although at a computational cost that might be significant. McDonald [1993] denotes that the NE context can be captured through internal evidences (as in the presence of "Sea" in "North Sea") and external evidences (as in "city of" before the NE "Seattle").

Palmer and Day [1997] describe a statistical approach for the NER task, where they report that basic heuristics work pretty well for some type of entities, and achieving high accuracy scores with their NER system for the given task. Mikheev et al. [1999] report a study on the importance of gazetteers for the task, concluding that regarding placenames, the gazetteer size is important.

NER systems typically fall in two categories:

1. Language-dependent systems, with manually crafted grammar rules that seize the knowledge required in each language to recognise entities in the text. These typically enforce strong NLP techniques such as morphosyntactic analysis. The use of gazetteers varies for each system, but nonetheless there are arguably stronger efforts on analysing text for evidence rather than looking up gazetteer entries. The NER system Palavras is an example of such systems [Bick, 2003].

2. Language-independent systems, which explore machine-learning techniques, thus having little or none human intervention on the whole NER process. The development of such NER systems is mostly restrained by the lack of annotated corpora to train their learning algorithms. For instance, the NERUA system [Ferrández et al., 2005] uses three machine learning classifiers (Hidden Markov Model, Maximum Entropy and Memory Based Learning) for the task, and based their participation on GeoCLEF in 2005 [Óscar Ferrández et al., 2006].

There are also NER approaches that combine both worlds, combining NLP techniques with semi-supervised learning methods (for instance, the work of Nadeau [2007]).

Recently, NER researchers started using Wikipedia on their systems, mostly for the NE disambiguation task. The works of Bunescu and Pasca [2006], Cucerzan [2007] and Kazama and Torisawa [2007], for instance, explore the Wikipedia categories, redirection pages, disambiguation pages, title structures and anchor texts to disambiguate NEs and train machine learning approaches to capture NER contextual information. Their approaches are based on the "one sense per discourse" assumption of Gale et al. [1992], where a given NE has a single context on the document. The surrounding NEs can help to determine such context by comparing key features on their corresponding Wikipedia pages (for instance, it is more likely that the NE "Armstrong" refers to a person if there is a NE "NASA" in the document, as the Wikipedia pages of "Neil Armstrong" and "NASA" are more closely related by links and categories than, for example, Wikipedia page the place "Armstrong (Ontario)").

Malin [2005], on the other hand, turns to social networks for NE disambiguation. The reported experiments on name disambiguation of actor collaborations from a snapshot of the Internet Movie Database (`http://www.imdb.com`) suggest that social networks provide a more robust environment for NE disambiguation than exact sources.

4.2 NER tools

The GATE framework [Cunningham et al., 2002] is an off-the-shelf tool that includes the ANNIE system. ANNIE (`http://gate.ac.uk/ie/annie.htm`) is an IE system

with basic NER capabilities optimised for English, which relies on a pipeline of finite state algorithms, including tokenisers / sentence splitters, gazetteer lookup modules, part-of-speech taggers and semantic taggers based on the Java Annotation Patterns Engine [Cunningham et al., 2000].

MITRE's Alembic Workbench system (**AWB**, `http://www.mitre.org/tech/alembic-workbench/`) is a natural language environment for developing tagged corpora. The AWB is capable to perform several IE tasks such as template filling or co-reference detection, and has also the capability of acquiring domain-specific tagging heuristics in an automatic way.

Baile (`http://balie.sourceforge.net/`) is also an IE framework that has NER capabilities, as shown in the YooName service (`http://www.yooname.com/`) [Nadeau, 2005]. Built over the Unstructured Information Management Architecture (UIMA) SDK and using the Weka toolkit for machine learning [Witten and Frank, 2005], Balie is able to recognise NEs for several languages, according to over 100 categories [Nadeau, 2007].

LingPipe (`http://alias-i.com/lingpipe/`) is a suite of NLP tools written in JAVA [Alias-i]. It includes modules for shallow tasks such as tokenisation, part-of-speech tagging or general chunking, as well as more advanced approaches for coreference resolution and NER. For the NER task, LingPipe implements a supervised training of a statistical model or, in alternative, a more direct method bases on dictionary lookup and regular expressions.

Palavras [Bick, 2000] is a constraint grammar (CG) framework for Portuguese, which performs high level morphosyntactic and semantic analysis. Based on the CG framework, grammar rules can be created to identify and disambiguate NEs [Bick, 2003]. A NER system based on Palavras achieved the highest performance scores in the first edition of HAREM [Bick, 2006].

NooJ [Silberztein, 2004] (`http://www.nooj4nlp.net/`) is a linguistic development environment with several tools for morphosyntactic analysis of texts. Mota and Silberztein [2007] developed a NER system over NooJ.

4.3 NER task

The task of NER was first coined in the 6th edition of MUC [Sundheim, 1995], which stipulated three major categories for NE classification (`ENAMEX`, `TIMEX` and `NUMEX`), encompassing entities referring to places/organizations/persons, to temporal expressions and to numeric expressions, respectively. In MUC, some issues, such as NER vagueness and semantic annotation according to NE context, were overlooked [Santos, 2007].

The MUC-6 evaluation methodology was based on the comparison of the annotations

made between the human annotators and the NER system, over a common set of documents (called the "gold standard" by Hirschman [1998], and later "golden collection" by Santos and Cardoso [2007]). the participants only had access to the non-annotated version of the gold standard (Santos and Cardoso [2007] even diluted such documents into a bigger collection, to better dissimulate it), which had to be automatically annotated by the NE system and returned to the organisers within a sort period of time. In MUC-6, participants had access to a manually NE-annotated document piece of training data, as there were a significant amount of machine-learning based NE systems, and MUC even organised some dry runs.

MUC-6 evaluation metrics were based on precision and recall measures, computed from the number of NEs that were correct, missing or spurious on the participant's gold standard documents [Douthat, 1998]. The F-measure was used to leverage the performance of all NER systems, and after a statistical significance analysis based on randomisation tests [Chinchor, 1992], MUC was able to compare NER approaches and point out which of them are best suited for the NER task.

After MUC, the CoNLL shared tasks revisited the NER task in 2002 and used a four-category classification (`PERSON`, `ORGANIZATION`, `PLACES` and `MISC`) on an already tokenised collection, using a similar NER task definition, methodology and evaluation measures [Sang, 2002, Sang and Meulder, 2003]. In 2004, the ACE contest [Doddington et al., 2004] notably had several improvements on the NER task definition, by using a wider range of categories and somehow supporting the notion of annotation in context on the NER task directives.

The HAREM NER contest took place in 2006, and redefined the NER task from a linguistic point of view by defining a categorization hierarchy comprising 10 main categories and 41 types, through careful analysis of sample documents made by several annotators [Santos and Cardoso, 2007]. The goal of HAREM was to define the NER task as seen by the research community and create a demanding task, regardless of the capabilities of the existing NER systems. Indeed, evaluation initiatives must sometimes enlight the path, leading the way and pointing out where researchers should place their efforts Hirschman [1998]. NER vagueness and annotation in context were a core subject on the HAREM task definition, directives and measure formulas, encouraging NER systems to capture the true meaning of each NE with respect to its context.

As such, HAREM collections comprise about 1200 documents from several textual genders (newspaper, fictional, political, etc), sources (web pages, mail, Wikipedia articles) and Portuguese variants, from which a subset of about 130 documents were selected as gold standards and were thoroughly annotated and reviewed by many annotators with HAREM categories and types, morphologic information (a morphological subtask was also included) and alternative identifications and classifications for vague/ambiguous NEs.

HAREM evaluation measures were an extension of the MUC measures, handling partially correct NEs and precise selection of alternative NE, proposing an combined classification score formula that summarises the identification, disambiguation and classification aspects of the NER task into a single measure value [Santos and Cardoso, 2007].

A detailed description of the first HAREM task is found in Santos and Cardoso [2007] and in Cardoso [2006]. In 2008, a second edition of HAREM [Santos et al., 2008] introduced minor corrections in the NER task.

5 Toponym Resolution

The toponym resolution (TR) task concerns on the mapping of already disambiguated placenames to their places. As Martins [2008] points out, knowing that a certain placename "Lisbon" mentions a place does not tell us which one, whether it refers to a city ("Lisbon, Portugal" or "Lisbon, ND, USA"?), a street or the metropolitan area, among other places that share the same placename. Note that the goals of TR do not concern whether such mapping is made through georeferencing on a geographic gazetteer or ontology or by geocoding with geographic footprints, but only on algorithms that approximate the human judgement when facing a given placename, and understanding the right place as mentioned in the text.

The PhD thesis of Leidner [2007] presents a comprehensive analysis of all related work on TR, namely the works of Hauptmann and Olligschlaeger [1999] about georeferencing speech transcriptions of broadcast news, the Perseus Project aimed to georeference a digital library of historical documents [Smith and Crane, 2001], the InfoXtract system which combines machine learning and hand-crafted rules for TR [Li et al., 2003], MetaCarta's approach on computing confidence estimations on each georeference [Rauch et al., 2003], the research of Pouliquen et al. [2004] around georeferencing placenames for multiple languages, the aforementioned Web-a-Where system [Amitay et al., 2004], the work of Schilder et al. [2004] around TR for documents in German, SPIRIT's approach on TR [Clough et al., 2004], the work of Zong et al. [2005] over assigning geoscopes for web documents, and the WikiDisambiguator system from [Overell and Rüger, 2006] which explores Wikipedia for placename grounding to TGN entries. Li et al. [2006] introduces a probabilistic approach for TR, based on the occurrence of placenames in the text and their hierarchical relationships as given by TGN.

Leidner [2007, pp. 111] presents a taxonomy of TR heuristics used by all of these works. In the taxonomy root node, heuristics are classified by their dependence on linguistic knowledge, world knowledge (unambiguous placenames are easily grounded without TR heuristics) or both.

5.1 Linguistic heuristics

TR heuristics that rely on linguistic knowledge use simple patterns such as "contained-in" patterns ($t_1,\ t_2$ or $t_1\ (t_2)$, as in "London, UK" or "London (UK)") or feature type patterns (as in *cityof* t_1 or t_2 *river*) to resolve a toponym. Some works take in consideration the placename frequency on the document to assign an importance weight.

5.2 World heuristics

TR heuristics based on world knowledge typically include population statistics (preferring the most populate places, or pruning gazetteer information for places with little population), feature type hierarchies (for instance, when a country is preferred over a city, or when a capital has always preference), resource authority (some gazetteers are more reliable than others), or even human intuition on having a default referent list for some placenames.

5.3 Linguistic and world heuristics

Some TR heuristics combine more complex NLP strategies with knowledge sources, such as gazetteers or ontologies. These heuristics try to go at the discourse level either by correlating the document subject to its placenames (for instance, finding related persons or events in the document may help contextualizing the placename reference), or by adopting simplistic approaches as the "one sense per discourse" coined by Gale et al. [1993], where it is assumed that ambiguous placenames have a common grounding context (for example, the presence of "London, UK" somewhere in a document implies that the remaining references to London are for the capital city), or through geometric minimality approaches. Geometric minimality heuristics use spatial distance between places to base toponym resolution, and typically aim to find a minimal bounding polygon. For instance, for a document that mentions Paris and Texas, a bounding polygon with t_1="Paris, Texas" and t_2="Texas" will be much smaller than one including t_1="Paris, Texas" and t_2="France", thus mimicking the human notion of geographic proximity.

5.4 Document TR

Also worth mentioning is the approach of Martins and Silva [2005] for assigning document scopes, proposing a graph-ranking algorithm that is concerned only with grounding a document, and not requiring the grounding of all placenames. Placenames such as "Lisbon", which may be associated to several feature types (city, municipality, district), have also strong *part-of* relationships in an ontology; so, a tree-like graph is created, where the nodes represent placenames (weighted by their frequency on the document), and are connected according to their ontological relationships. After applying an iteration approach inspired by the PageRank algorithm [Page et al., 1998], a decision is made on the most representative node for each document, which is subsequently georeferenced to the geographic concept of the ontology.

6 Geographic indexing and geo-similarity

Indexing focuses on representing information extracted from document collections into optimised data structures, allowing fast responses by retrieval modules. *Ranking* concerns on strategies to compute similarity measures that mimic the human notion of relevance between documents and queries regarding a given subject (time, space, subject, authority, etc). To compute relevance, one should define an n-dimension *space model* where documents and queries are represented, and the retrieval process is performed by algebraic algorithms. *Weighting* algorithms provide formulas to compute similarity measures between documents and queries, in the used space model.

Most indexing and ranking approaches are often reported as a single step, as index structures are often designed to suit the requirements of ranking and weighting approaches. For instance, if the ranking algorithm gives more importance to terms in the first sentences of the document and/or the weighting scheme needs document frequency statistics to know the rarity of terms, the indexing structure should encode information about term position in the document and document frequency for each term.

Woodruff and Plaunt [1994] already mentioned the need for an automatic indexing step for text documents in terms of geographic location, so that the text could be integrated with georeferenced data. Challenges in geographic indexing and retrieval include the design of efficient data structures for geographic metadata (from geometric shapes to ontology identifies), and the development of geographic weighting schemes that approximate geo-similarity, as understood by humans. Martins et al. [2005] overview some work on geographic indexing and ranking.

6.1 Term indexing and ranking

Term indexing goes way back to the work of [Luhn, 1957], which proposed terms as units for indexing, which could be used as selection criteria. The first term indexes used only terms from manually-assigned categories or titles, but nowadays it is a common practice to index whole documents. Term indexes are widely used in IR, as they suffice in most cases [Witten et al., 1994]. Some optimizations and improvements of term indexing include stop-word removal, lemmatization or stemming (which may lead to some information loss).

6.2 Term weighting

Term weighting schemes approximate a relative importance of different terms for a given retrieval. In the Vector Space Model (VSM) proposed by Salton et al. [1975], vectors are

used to describe term occurrences of documents and queries in a n-dimensional vector space. The vector length is therefore used as a basic term weighting measure, often linked to the term frequency ($tf_{t,D}$), or the number of occurrences of a term t in a document D, following the heuristic that a document is likely to be more relevant for a given query if it contains more instances of query terms.

Spärck Jones [1972] proposed a term weighting scheme based not only on tf, but also on the inverse document frequency (idf_t), following also the heuristic that more discriminative terms occur less in a document collection [Spärck Jones, 2004]: the $tf - idf$. As summarised by Efthimiadis [1993], "very frequent terms are not very useful, middle frequency terms are quite useful, infrequent terms are likely to be useful but not as much as the middle frequency terms, and very infrequent terms are useful terms in the sense that when they are present are good indicators; however, since these terms are not present for most of the time they do not help in retrieving very many documents."

The $tf - idf$ was widely used to represent term vectors in the SVM, and along with the *cosine similarity* measure, which computes similarity between two vectors in a [-1, 1] range as given by the angle of the cosine, it was the core indexing and ranking algorithm of almost all IR systems in the 70s and 80s. A slight modification of the $tf - idf$ with a logarithmic value of tf was found to produce better results with the SMART system [Buckley et al., 1992]. The SMART retrieval system was a great platform to experiment several text weighting approaches, such as pivoted normalization weights, where document length (dl) values were also used to damp the fact that longer documents contain more terms, thus skewing the tf values and penalizing short but concise documents [Salton, 1971].

The works of Harter [1974] and Bookstein and Swanson [1975] concerning the Poisson distribution and term distribution were the ground base for the Okapi experiments and the development of the BM weighting schemes [Robertson et al., 1992, Robertson and Walker, 1994], based on probabilistic models. The BM-25 version of such term weighting scheme [Robertson et al., 1994] is one of the state-of-the-art text weighting schemes on IR, and it is widely used in the community.

6.3 Geographic indexing

While GIS is all about geographic information put into data structures, such structures are mostly focused on precise representation, rather than retrieval efficiency. Conversely, a GIR system does not need detailed polygons of geographic footprints; simpler shapes will suffice, as retrieval should be efficient to avoid being the system bottleneck.

An overview of multi-dimensional indexes can be found in Gaede and Günther [1998], mostly based on grid and tree structures. Grid structures are simple representations of

footprints using area units from a fixed grid. For instance, the SPIRIT project uses a grid scheme to index geographic footprints [Jones et al., 2004]. Vaid et al. [2005] experimented with different indexing strategies for text and spatial data, to determine the best indexing strategy for for geographic retrieval.

Among the several tree structures, R-tree is well suited to the GIR task as it allows an easy representation of overlapping rectangles, it can be balanced and optimised to speed up access to its information, and it allows basic geographic operations such as intersection or nearest neighbors [Martins et al., 2005]. The R*-trees [Beckmann et al., 1990] is an improvement over the R-tree by incorporating a combined optimization of area, margin and overlap of each enclosing rectangle in the directory, with a slight implementation cost. Zhou et al. [2005] propose hybrid index structures that integrate inverted term indexes with R*-trees, and also conclude that R*-trees are more efficient than grid-based indexes for geographic data.

For GIR systems that georeference documents into ontologies or gazetteer entries rather than geographic footprints, geographic indexing can be achieved by simply adapting inverted term indexes to the unique geographic identifiers, where a single access can therefore return all documents that are associated with a certain geographic concept.

6.4 Geo-similarity

in the early days, as most GIR prototypes were in fact proof-of-concepts and not fully-working retrieval systems, topics such as geographic indexing and geo-similarity were normally unaddressed. In fact, a good approach for measuring geo-similarity is an important component on the document ranking strategy of a GIR system.

Jones et al. [2001] first proposed the use of ontologies to generate distance measures to be used as geo-similarity measures for document ranking, an approach that was subsequently adapted in the GIR prototype of the SPIRIT project. A simple weighted combination of spacial distance measures and thematic measures was used to give a final relevance score for document ranking (Cai [2002] proposed a similar approach for his GeoVSM model).

MetaCarta [Rauch et al., 2003] uses a proprietary technology called CartaTrees^{TM} to index and compute geo-similarity. While the details of the overall ranking algorithm are not available, it is known that it uses geographic references and coordinates, along with the full text of each document, to produce text indexes and spatial indexes.

The SPIRIT project addressed the problem of geo-similarity by computing Euclidean distances between document and query footprints Vaid et al. [2005]. The term similarity score (given by a BM25 weighting scheme) and geo-similarity score (given by footprint distances) are then combined through a distributed relevance ranking approach proposed

by van Kreveld et al. [2004]. This document ranking approach addresses the topic of *spatial diversity*, where users querying about a given subject within a geographic scope, with the purpose of knowing more about such subject, are likely to be interested in documents that are geographically scattered around the query scope, rather than reading several documents over a smaller, focused area.

Markowetz et al. [2005] summarises the dilemma of using geographic metadata to compute geo-similarity scores for document ranking in two distinct methods: i) use the query footprint as a filter, thus selecting only the documents that have a geographic footprint within the query area (and still use only term weighting scores for document ranking), or ii) compute a geo-similarity score that represents the geographic proximity between document and query footprints, and combine it with term weighting score.

The GReaSE project adopted a group of heuristic approaches to compute geo-similarity between scopes, relying on geographic reasoning in an ontology [Andrade and Silva, 2006]. These heuristics are: i) topological distance, as in "London" and "UK", given by node distances in the ontology, ii) spatial distance, using bounding boxes to compute distances between scopes; a double sigmoid function was used to smooth the relative notion of distance according to the scopes sizes [Egenhofer and Mark, 1995] (that is, Norway is near Sweden but Oslo is not near Stockholm), iii) shared population and iv) adjacency, given by ontologic relationships [Martins et al., 2007].

7 Geographic query reformulation

Query reformulation (QR) is the process of enhancing an initial query string, with the purpose of building a final query that better reflects the initial information need, and better takes advantage of the underlying search engine capabilities (such as the case of GIR systems for geographic queries). In the literature, query reformulation is more commonly referred to as *query expansion* (QE), because the most widely used technique for query reformulation is the addition of strongly related terms to the initial query. To emphasise this distinction, we will keep addressing the whole query enhancement process as query reformulation, and so the reader should be aware of this terminological difference between other works and this survey, namely the referential surveys on query expansion made by Efthimiadis [1996] and Ruthven and Lalmas [2003]. *Query re-weighting* is another QR technique, which assigns importance scores to query terms according to their relevance on subject description. Note that query expansion and query reweighting are two different activities, as pointed by Harman [1992] and Harman [1988].

Geographic query reformulation can be seen as a specialization of the query reformulation task devoted for query strings that have a geographic scope. Geographic QR is a two-fold task: i) it improves the clarity of the addressed subjects and mitigate the terminological differences, as in standard QR, and ii) it redefines the geographic criteria of the query into a group of explicit geographic concepts, according to the spatial relationship of the query. For instance, a query "Tourism in Portuguese islands" should be expanded to address both touristically-related terms such as hotels, beaches or sightseeing tours, and clear out the geographic entities of the query scope, such as expanding to the names of all islands from the Portuguese territory, or generating its footprint.

There are few published works dedicated to the specific problems of geographic query reformulation, although there are several reports on the analysis of geographic queries. In fact, one of the main goals for the proposed PhD work of the author is precisely the development of a new QR module focused on GIR that employs semantic-flavored techniques to address the problems of geographic QR.

Regarding previous work in geographic QR, Fu et al. [2005] used an ontology-driven QR module that derives the query footprint and resolves directional queries. [Cardoso and Silva, 2007] performed geographic QR with different expansion strategies according to the feature types and spatial relationships addressed in the query. Delboni et al. [2007] proposed a semantic-based QR strategy to identify placenames and spatial relationships, generating final queries that were afterwards submitted to Google. [Stokes et al., 2008], on a study about the impact of the performance of each GIR task on the overall GIR performance, proposed a new QR framework devoted to minimise the errors derived from GIR components. Their work addressed the problem of *query overloading*, that is,

when the geographic QR approach generates significantly more geographic terms than the standard term expansion, biasing the document retrieval towards geographically-related documents but with unrelated subjects.

This chapter starts with an overview of the geographic queries, and then a detailed description of the state-of-the-art in query reformulation.

7.1 Characterizing geographic queries

A typical approach for handling geographic queries is to assume that they can be split into $<$ *what*, *spatial relationship*, *where* $>$ triplets) [Jones et al., 2004, Martins et al., 2007], and subsequently handle each field separately (although there are doubts concerning whether this approach is a good practice or not [Cardoso and Santos, 2008]). Detecting query types is also important to determine when queries have a geographic scope of interest. For instance, a query for "France Press" is not a geographic query but rather likely a navigational query, although the term "France" might be misleading. Gravano et al. [2003] addressed this subject, by pointing out that some queries are better satisfied by documents from a local scope (as in a city), while others are from a global scope (as in a country). Their experiments with several machine-learning classification approaches of geographic queries show that retrieval performance can be greatly improved if query locality can be accurately detected, and suggest that QR approaches use such information.

Kohler [2003] analysed the query logs submitted by users in a single day from the Excite search engine. Roughly one fifth of the queries can be considered of geographic interest, and of these nearly 80% had a placename. It is also reported that many spatial relationships such as "south" or "near" were used when the geographic criteria is not the typical inclusion denoted by "in" or "at", and that there were many feature types such as "county" and "city" among the highest frequently geographic terms list. Gan et al. [2008] recently made a more detailed analysis of geographic queries for the query log corpus released by AOL search engine Pass et al. [2006].

Within the SPIRIT project, Vaid et al. [2005] classified geographic queries according to their spatial relationship: i) *proximal*, where a certain distance is specified with exact distances, as in "*schools within 10 km of Zurich*", or fuzzy distances, as in "hotels near Cardiff University "; ii) *topological*, where an overlapping or adjacency criteria is used, as in "hospitals in London"; and iii) *directional*, which specifies a point of start and a given orientation, as in "*holiday resorts north of Milan*". These classes are quite similar compared to the formalization made by Pullar and Egenhofer [1988] in: i) topological relationships, ii) spatial order and iii) metric relationships (although the latter was not focused on GIR).

Following the initial classification of geographic topics by Santos and Chaves [2006],

GeoCLEF organisers suggested the following classification for geographic topics [Gey et al., 2007a]:

1. Non-geographic subject restricted to a place, as in "Shark Attacks off Australia and California";
2. Geographic subject with non-geographic restriction, as in "Cities near active volcanoes";
3. Geographic subject restricted to a place, as in "Cities along the Danube and the Rhine";
4. Non-geographic subject associated to a place, as in "Independence movement in Quebec";
5. Non-geographic subject that is a complex function of a place, as in "Water quality at the coast of the Mediterranean";
6. Geographic relations among places, as in "How are the Himalayas related to Nepal?";
7. Geographic relations among (places associated to) events, as in "Did Waterloo occur more north than Agincourt?";
8. Relations between events requiring their precise localisation, as in "Was it the same river that flooded last year and in which killings occured in the XVth Century?".

As stated in Mandl et al. [2008b], GeoCLEF topics in 2006 and 2007 were exclusively from kinds 1 and 2, while in Mandl et al. [2008a], although no classification is provided, from the report it is clear that the same procedure (and therefore kind of topics) was used. Overell [2009, pp. 147] also denotes that the most frequent geographic topic in GeoCLEF is by far the topics about non-geographic subject restricted to a place. Topics also include several challenges for detecting the query scope, such as vague/ambiguous places ("St. Paul's Cathedral", "Eastern Bloc"), different placenames and language variants ("Myanmar / former Birmania", "Lisbon / Lisboa / Lissabon"), several granularity levels and feature types (from continents to cities, from volcanoes to water bodies), and query scopes bounded to a single area ("Europe") or to multiple areas ("Cities near active volcanoes").

7.2 QR overview

QR is a widely used technique to improve search results of IR systems. For example, Braschler and Peters [2004] revealed that the best IR systems that participated in the CLEF 2002 campaign relied on robust stemming, a good term weighting scheme and a query reformulation approach. The hypothesis behind QR is that by reformulating the initial query with additional related terms, we increase the odds of matching terms of relevant documents [Xu and Croft, 1996]. QR can be seen as an approach to assist the IR system to capture the concepts from the query terms, narrowing the terminological gap created by the different vocabulary used by users' queries and authors' documents to express the same concepts.

QR is reported to significantly improve the quality of search results [Salton and Buckley, 1988, Lu and Keefer, 1995, Xu and Croft, 1996], but not for all queries; in fact, QR sometimes worsens the retrieval results. There are many reasons for this:

1. Short and imprecise queries are harder to expand since the underlying information need of the user is vague. Such short queries account for the majority of searches in web search engines [Jansen et al., 2000].

2. Failing to understand the true meaning of the query will lead to QR drift from the original query [Mitra et al., 1998].

3. Most QR have parametrised approaches, and such parameters are not the optimal parameters for all kinds of queries [Billerbeck and Zobel, 2003].

4. Users do not spend much time formulating query strings and rarely use the logic operators available on most search engines such as Boolean operators or exact phrase searches, and query structure has its own share of influence on retrieval results [Järvelin et al., 1998].

QR is a technique that has its benefits for IR when properly implemented according to the type of queries and document collections used. For instance, Billerbeck [2005] showed that the same QR approach produces different results, just by switching the document collection from newspaper collections to web collections. Also, the utility of QR depends on the query type; for instance, a user querying for "apple" might just be searching for the website address of Apple Inc. (`www.apple.com`), and thus QR should not be used in this case.

Broder [2002] classified the underlying user needs in three types: navigational (as in the above example, where user is searching for a named page), transactional (looking for a given service to buy products or download files) or informational (searching for

additional information on a given topic). Aires [2005] also presented similar conclusions on an empirical study made to classify user needs over the Portuguese web.

Additionally, while QR is important for vague queries, it is not so useful for well-defined queries. Cronen-Townsend and Croft [2002] addressed this problem by proposing a *clarity score* that quantifies the vagueness and/or ambiguity level on the query (see Santos [2004] for more information about the differences between vagueness and ambiguity). The clarity score is measured by comparing the language models of both query and collection, and can be used to estimate the query performance on a subsequent retrieval [Cronen-Townsend et al., 2002]. The clarity score can be used by QR to adjust its parameter configuration, and somewhat control the amount of query modifications for each query.

7.3 QR approaches

While in QR there are countless strategies for term selection and weighting (normally fitting the type of source / collection used and the retrieval purposes), they somehow try to replicate the query reformulation actions normally used by users, when they are faced with search results that do not satisfy their intentions (*manual query reformulation*). User techniques for query reformulation vary greatly, but they normally include: i) addition of more terms, either to clarify an information need (for instance, adding "animal" to a query "jaguar" to specify that one is only interested in results about the animal), or to fine-grain the results; ii) remove terms, normally used when the search results do not present enough relevant results, and thus performing a broader search, or iii) changing terms, for instance when the user is searching for a given subject but he is not able to express it with the most adequate terms, but after several searches, the user tries out new terms learnt from past search results. **?** analysed the reformulation strategies made by users that submitted geographic queries to the Yahoo! search engine.

QR approaches can be divided as: i) automatic query reformulation (AQR), for QR strategies with no human intervention on the reformulation of the query string, and ii) interactive query reformulation (IQR, sometimes called semi-automatic query reformulation), for QR strategies where there is any kind of human input that influences the final query.

7.3.1 QR based on search results

The first works on QR date from the 1960s with the proposal of Rocchio Jr. [1971] (and later refined by Ide [1971]) for a *relevance feedback* (RF) technique based on the search results, which were tested in the SMART system. RF is grounded on the assumption that, if search results contain relevant documents on the top for a given query, these documents

might have other strongly related terms that might be used to enhance the query. In the same line of thought, documents ranked in the bottom of the results list are likely to have non-related terms that also might be used as negative examples for term selection and weighting. Rocchio's formula can be iterated several times, generating an optimal final query which ideally separates all relevant documents from the non-relevant documents.

Robertson and Spärck Jones [1976] proposed the binary independence model (BIM), a probabilistic model for the weighting of additional terms, based on the distribution of terms among relevant and non-relevant documents (hence the binary reference). The model was later refined to the F4 formula [Robertson, 1986], and furthermore by Efthimiadis [1993] which combined Robertson's formula to propose the $w_t(p_t - q_t)$ algorithm, according to the ideas transmitted by Robertson [1990] on which the selection of new terms and their re-weighting require different weighting formulae. The $w_t(p_t - q_t)$ algorithm achieved better results in the context of an experiment over interactive query reformulation [Efthimiadis, 1993].

Salton and Buckley [1990] compared the relevance feedback algorithms over several collections, and reported that Ide's RF approach produces slightly better results. On the other hand, Harman [1988, 1992] focused in comparing the impact of query expansion and term re-weighting on the overall RF approaches, and concluded that the query expansion step is more relevant, and that the best results are achieved when a limited number of terms are used in QE (around 20 terms). Harman also reported that multiple iterations on RF are highly effective.

As such, a fully automatic QR based on relevance feedback has to make two decisions: i) estimate the number of top documents that are relevant for the given query (*top-k-docs*), and ii) select a limited number of terms for query expansion (top-k-terms). These threshold values are normally pre-defined, and such RF approaches are called *pseudo relevance feedback* (or *blind relevance feedback*), see Buckley et al. [1992]. In his work, Mitra et al. [1998] showed that it is important to ensure that the documents above the *top-k-doc* threshold are indeed relevant; his experiments with re-reranking of documents before QE using refinement heuristics based on proximity constraints and fuzzy Boolean operators showed significant improvements in the RF step.

Another interesting work around RF was made by Haines and Croft [1993], who extended the inference networks proposed by Turtle [1990] to include RF techniques and reported significant improvements on retrieval performance. The work of Yang and Korfhage [1994] adapted genetic algorithms to RF, using a fitness formula that compares the effectiveness of a query against a defined set of documents known to be relevant. An overview of genetic algorithms for RF is given by López-Pujalte et al. [2002].

7.3.2 QR based on past user queries

In interactive QR, the user can intervene on the key steps of QR, by supervising query expansion (for instance, choosing terms from the top-k-term subset or redefining Boolean operators to cluster related terms) or by providing relevance feedback over documents (that is, selecting the relevant documents on which RF approaches should give positive feedback on term selection and re-weighting). Harman [1988] performed a set of experiments aiming to find the best techniques for term selection, involving RF, a nearest neighbor approach and term variants, where user interaction was simulated to select the top 20 query terms for the QE step.

As the user is often unwilling to spend time on giving feedback and selecting terms, such interactive QR strategies are rarely used in IR systems. Google recently launched the SearchWiki feature (`http://googleblog.blogspot.com/2008/11/searchwiki-make-search-your-own.html`), where each user can interact with the results list (by promoting, demoting or deleting documents) and thus provide feedback for further searches.

In any case, research work around IQR is more promising focusing on the analysis of the implicit interactions made by users, as recorded in search engine query logs. Assuming that users, after issuing a query q, analyse the search results and choose to visit a given document d (either because of its title, website or the automatically generated summaries, called *snippets*), they are implicitly saying that they found that document d is relevant to query q. Fitzpatrick and Dent [1997] adapted the RF approach to use query logs as a feedback source, reporting interesting results; Cui et al. [2002] selected terms for QE by exploring term co-occurrence between queries that were issued by several users that selected the same documents. Anick [2003] closely studied the manual query reformulations of past users, and analysed the refinement effectiveness achieved on such user sessions. Baeza-Yates [2004] and Baeza-Yates et al. [2004] explored query logs to generate query recommendations for users. Kraft and Zien [2004] explored anchor text as a source of query terms for QE. [Fonseca et al., 2005] focused their work on capturing the concepts behind user queries, by building a graph of query relations using association rules. Agichtein and Zheng [2006] used query logs to train machine learning approaches to choose the most relevant documents for the most frequent queried subjects.

7.3.3 Other QR sources

Other than the aforementioned search results and query logs, there are other information sources that have been used by QR to base their expansion strategies. Early research on QR used thesaurus to find synonyms and cluster terms [Spärck Jones, 1971]. While the use of thesauri seems a natural choice to mitigate the terminological differences between queries

and documents, thesauri-based QR approaches failed to present significant improvements in retrieval results [Voorhees, 1994].

Qiu and Frei [1993] used a similarity thesaurus built automatically from documents (thus restraining the QE action to the subjects addressed on the collection), and adding terms that are similar to the query concept, not the individual query terms. A probabilistic QR module using this thesaurus presented a significative improvement on the results, showing that thesauri-based QR should are valid approaches, as long as they use thesauri fit for the QR task.

Xu and Croft [1996] compared the performance between two distinct trends on QR approaches: i) global QR techniques, which explore the whole document collection for additional terms and co-occurrence data, and ii) local feedback QR techniques, which explored search results as addressed above. While local feedback QR approaches outperformed global QR approaches, the best results were achieved by the proposed local context analysis approach [Xu and Croft, 2000], which combined both collection and search result data.

Carpineto et al. [2001] proposed an information theoretic approach for QR, by analysing the divergences between term distributions over the whole collection, and on the (pseudo) relevant documents from local feedback. Carpineto's distribution analysis approach was used to weight terms on Rocchio's formula, and reported improvements on retrieval performance.

Sanderson and Croft [1999] proposed an automatic way to extract concepts from large collections, by applying shallow parsing strategies to find term patterns like "X such as Y, Z" or "X, Y and other Z" build an hierarchy of concepts. The hierarchy could be used by QR approaches to perform focused reformulations around specific concepts, thus avoiding, in a way, term ambiguity.

Billerbeck and Zobel [2005] made an interesting experiment, performing term expansion on the document side (document expansion), and compared the performance gain against QR approaches. The experiment aimed to automatically add related terms to documents and index such augmented versions, thus leaving the burden of term expansion to an offline process and relieving the online search process of such task. Nonetheless, document expansion showed little improvements over query expansion. Li et al. [2006] used document expansion just for grounded placenames, which also reduced the time required for QR. For instance, the placename "Victoria" in a document is expanded to "Victoria|Australia|Oceania" and indexed as such. For queries such as "Hotels in Australia", such document will be geographically ranked higher, without requiring complex comparisons between document and query scopes.

8 Conclusion

GIR is nowadays a well-established research field, playing a leading role on today's IR challenges on developing search engines that are more receptive to the real user's needs. Current research works on GIR report interesting advances on document geoparsing, new ways to explore GIR resources, dedicated GIR retrieval interfaces and personalised results according to geographic criteria (given by the user or by a GPS location of the mobile device), and in such fertile environment, we can only expect great outcomes in GIR on the next years.

As the vision of the Semantic Web and the Linked Data becomes more and more a reality, GIR research is a clear example on how it can greatly benefit from recent machine-dedicated data resources such as DBpedia and Geonames.org, and use it to develop human-friendly web applications. One can only expect that, as such data resources mature even more, GIR research can report more staggering advances towards its goal.

Nonetheless, since the GeoCLEF last edition in 2008, there is an alarming lack of evaluation initiatives for GIR systems and their components: today, only the GikiCLEF 2009 track will keep a geographic motivation on their evaluation task. Since evaluation is a key activity on any scientific, technological or engineering work, GIR research will now proceed without a community-based evaluation benchmark to compare different approaches and measure performance gains and failures on each GIR components. While the past GeoCLEF evaluation scenarios can still be used, they will no longer pose to the GIR community with new and challenging topics that will keep up (and lead out) with the state-of-the-art GIR systems.

Likewise, specific GIR tasks such as named-entity recognition or toponym resolution do not have a periodic evaluation contents, although there are suitable proposed methodologies to evaluate such tasks. This will make it harder to evaluate the performance bottlenecks of GIR systems, and to select the right tasks where research efforts must be focused on.

Finally, a remark on the NLP approaches on IR-related fields. While there has been a significant number of attempts of applying NPL techniques to improve IR results without encouraging results (Smeaton [1997] overviews this lack of progress on NLP-based IR systems), we feel that the time is ripe to try again and develop retrieval systems that try to understand the messages contained in queries and documents. There are a number of reasons why we believe that now NLP techniques can be useful for text retrieval systems:

- NLP-based approaches are normally connotated with more CPU-demanding processes, while simple statistic-based IR approaches are more scalable and have a more resilient performance rate with larger document collections [Brants, 2003]. Nonetheless, CPU cost is becoming more a cheaper commodity, and with the advent

of new grid-computing architectures and parallel-computing platforms such as Map Reduce, NLP-based approaches are now feasible for online retrieval systems.

- One of the main features of Web 2.0 is the changing role of users to active participants on the web data. Community-based resources, such as Wikipedia or Flickr, provide a dedicated interface to invite the user to easily edit metadata and structure their information. While such structured data may be encoded from tags to HTML elements, with dedicated middleware projects that generate RDF linked data from such information makes it much more easier to develop reasoning applications that to browse all the human knowledge without the need of supplementary NLP wrappers such as tokenisers or taggers.

- As the user is becoming more accustomed to use the Internet for its daily needs, either on generic information or in web services, there is a great opportunity on developing systems that can understand his needs, guide them to the particular topic or service that he is looking for, and present the results with a personalised page that suits each user profile and the context of each search.

References

ADL. Alexandria Digital Library Gazetteer. Santa Barbara CA: Map and Imagery Lab, Davidson Library, University of California, Santa Barbara. Copyright UC Regents., 1999.

Eugene Agichtein and Luis Gravano. Snowball: Extracting Relations from Large Plain-Text Collections. In *Proceedings of the 5th ACM Conference on Digital Libraries (DL'00)*, pages 85–94, San Antonio, TX, USA, June 2-7 2000. ACM.

Eugene Agichtein and Zijian Zheng. Identifying "Best Bet" Web Search Results by Mining Past User Behavior. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'2006)*, pages 902–908, Philadelphia, PA, USA, August 20–23 2006. ACM.

Rachel Aires. *Uso de marcadores estilísticos para a busca na Web em português*. PhD thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Agosto 2005. in Portuguese.

Alias-i. LingPipe 3.7.0. `http://alias-i.com/lingpipe`, 2008.

James Allan, Jay Aslam, Nicholas Belkin, Chris Buckley, Jamie Callan, Bruce Croft, Sue Dumais, Norbert Fuhr, Donna Harman, David J. Harper, Djoerd Hiemstra, Thomas Hofmann, Eduard Hovy, Wessel Kraaij, John Lafferty, Victor Lavrenko, David Lewis, Liz Liddy, R. Manmatha, Andrew McCallum, Jay Ponte, John Prager, Dragomir Radev, Philip Resnik, Stephen Robertson, Roni Rosenfeld, Salim Roukos, Mark Sanderson, Rich Schwartz, Amit Singhal, Alan Smeaton, Howard Turtle, Ellen Voorhees, Ralph Weischedel, Jinxi Xu, and ChengXiang Zhai. Challenges in Information Retrieval and Language Modeling: Report of a Workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts, Amherst, September 2002. *SIGIR Forum*, pages 31–47, 2003.

Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffe. Web-a-Where: Geotagging Web Content. In Mark Sanderson, Kalervo Järvelin, James Allan, and Peter Bruza, editors, *Proceedings of the 27th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'04)*, pages 273–280, Sheffield, UK, July 25–29 2004.

Leonardo Andrade and Mário J. Silva. Relevance Ranking for Geographic IR. In Ross Purves and Chris Jones, editors, *Proceedings of the 3rd ACM Workshop On Geographic*

Information Retrieval, GIR 2006, Seattle, WA, USA, August 10, 2006. Department of Geography, University of Zurich, 2006.

Peter Anick. Using Terminological Feedback for Web Search Refinement: a Log-Based Study. In *Proceedings of the 26th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'03)*, pages 88–95, Toronto, Canada, July 28 - August 1 2003. ACM.

Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, and Sriram Raghavan. Searching the Web. *ACM Transactions on Internet Technology*, 1(1):2–43, August 2001.

Sören Auer and Jens Lehmann. What Have Innsbruck and Leipzig in Common? Extracting Semantics from Wiki Content. In Enrico Franconi, Michael Kifer, and Wolfgang May, editors, *The Semantic Web: Research and Applications, 4th European Semantic Web Conference, ESWC 2007, Innsbruck, Austria, June 3-7, Proceedings*, volume 4519 of *LNCS*. Springer, 2007.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A Nucleus for a Web of Open Data. In Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11–15, 2007, Proceedings*, number 4825 in LNCS, pages 722–735. Springer, 2007.

Amittai E. Axelrod. On Building a High Performance Gazetteer Database. In András Kornai and Beth Sundheim, editors, *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, pages 63–68, Edmonton, Canada, May 31 2003. ACL.

Ricardo Baeza-Yates. Web Usage Mining in Search Engines. In Anthony Scime, editor, *Web Mining: Applications and Techniques*, pages 307–321. Idea Group, 2004.

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison Wesley, 1999.

Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. Query Recommendation Using Query Logs in Search Engines. In Wolfgang Lindner, Marco Mesiti, Can Türker, Yannis Tzitzikas, and Athena Vakali, editors, *Current Trends in Database Technology - EDBT 2004 Workshops, EDBT 2004 Workshops PhD, DataX, PIM, P2P&DB, and*

ClustWeb, Heraklion, Crete, Greece, March 14-18, 2004, Revised Selected Papers, volume 3268 of *LNCS*, pages 588–596. Springer, 2004.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open Information Extraction from the Web. In Manuela M. Veloso, editor, *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2670–2676, Hyderabad, India, January 6-12 2007.

Norbert Beckmann, Hans-Peter Kriegel, Ralf Schneider, and Bernhard Seeger. The R*-Tree: An Efficient and Robust Access Method for Points and Rectangles. In Hector Garcia-Molina and Hosagrahar Visvesvaraya Jagadish, editors, *Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data (SIGMOD'90)*, pages 322–331, Atlantic City, NJ, USA, May 23-25 1990. ACM Press.

Brandon Bennett and Pragya Agarwal. Semantic Categories Underlying the Meaning of 'Place'. In Stephan Winter, Matt Duckham, Lars Kulik, and Benjamin Kuipers, editors, *Spatial Information Theory, 8th International Conference, COSIT 2007, Melbourne, Australia, September 19-23, 2007, Proceedings*, volume 4736 of *LNCS*, pages 78–95. Springer, 2007.

Tim Berners-Lee. `http://www.w3.org/DesignIssues/LinkedData.html`, 2007.

Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, May 2001.

Eckhard Bick. Multi-level NER for Portuguese in a CG Framework. In Nuno J. Mamede, Jorge Baptista, Isabel Trancoso, and Maria das Graças Volpe Nunes, editors, *Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003, Faro, Portugal, June 26-27, 2003. Proceedings*, volume 2721 of *LNCS*, pages 118—-125. Springer, 2003.

Eckhard Bick. Functional Aspects in Portuguese NER. In Renata Vieira, Paulo Quaresma, Maria das Graças Volpe Nunes, Nuno J. Mamede, Claudia Oliveira, and Maria Carmelita Dias, editors, *Computational Processing of the Portuguese Language, 7th International Workshop, PROPOR 2006, Itatiaia, Brazil, May 13-17, 2006, Proceedings*, volume 3960 of *LNCS*, pages 80–89. Springer, 2006.

Eckhard Bick. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. PhD thesis, University of Aarhus, Aarhus, Denmark, November 2000.

Bodo Billerbeck. *Efficient Query Expansion*. PhD thesis, RMIT University, Melbourne, Australia, September 2005.

Bodo Billerbeck and Justin Zobel. When Query Expansion Fails. In *Proceedings of the 26th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'03)*, pages 387–388, Toronto, Canada, July 28 - August 1 2003. ACM.

Bodo Billerbeck and Justin Zobel. Document Expansion versus Query Expansion for Ad-hoc Retrieval. In *Proceedings of the 10th Australasian Document Computing Symposium (ADCS'2005)*, pages 34–41, Sydney, Australia, December 12 2005.

Susanne Boll, Christopher Jones, Eric Kansa, Puneet Kishor, Mor Naaman, Ross Purves, Arno Scharl, and Erik Wilde. Location and the Web: (LocWeb 2008). In *Proceedings of the 1st International Workshop on Location and the Web, LocWeb 2008, Beijing, China, April 22, 2008*, volume 300 of *ACM International Conference Proceeding Series*. ACM, 2008.

Abraham Bookstein and Don R. Swanson. A Decision Theoretic Foundation for Indexing. *Journal of the American Society for Information Science*, XXVI:45–50, January 1975.

Thorsten Brants. Natural Language Processing in Information Retrieval. In Bart Decadt, Véronique Hoste, and Guy De Pauw, editors, *Computational Linguistics in the Netherlands, CLIN 2003, December 19, Centre for Dutch Language and Speech, University of Antwerp*, volume 111 of *Antwerp Papers in Linguistics*. University of Antwerp, 2003.

Martin Braschler and Carol Peters. Cross-Language Evaluation Forum: Objectives, Results, Achievements. *Information Retrieval*, 7(1-2):7–31, 2004.

Andrei Broder. A Taxonomy of Web Search. *ACM SIGIR Forum*, 36(2):3–10, 2002.

Chris Buckley, Gerard Salton, and James Allan. Automatic Retrieval with Locality Information Using SMART. In *Proceedings of the 1st Text REtrieval Conference (TREC-1)*, pages 59—-72, Gaithersburg, MD, USA, November 1992. NIST Special Publication 500-207.

Razvan Bunescu and Marius Pasca. Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 9–16, April 3-7 2006.

Davide Buscaldi and Paolo Rosso. On the Relative Importance of Toponyms in Geoclef. In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, Anselmo Peñas, Vivian Petras, and Diana Santos, editors, *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, volume 5152 of *LNCS*, pages 815–822. Springer, 2008a.

Davide Buscaldi and Paolo Rosso. Geo-WordNet: Automatic Georeferencing of Wordnet. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC'08)*, Marrakech, Morocco, May 28-30 2008b.

Davide Buscaldi and Paolo Rosso. The UPV at GeoCLEF 2008: the GeoWorSE System. In Carol Peters et al., editors, *Working Notes of CLEF 2008*, Aarhus, Denmark, September 17-19 2007.

Davide Buscaldi, Paolo Rosso, and Emilio Sanchis Arnal. Using the WordNet Ontology in the GeoCLEF Geographical Information Retrieval Task. In Carol Peters, Fredric C. Gey, Julio Gonzalo, Henning Müller, Gareth J. F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke, editors, *Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evalution Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers*, volume 4022 of *LNCS*, pages 939–946. Springer, 2006.

Orkut Buyukkokten, Junghoo Cho, Hector Garcia-Molina, Luis Gravano, and Narayanan Shivakumar. Exploiting Geographical Location Information of Web Pages. In *ACM SIGMOD Workshop on The Web and Databases (WebDB'99)*, pages 91–96, 1999.

Guoray Cai. GeoVSM: An Integrated Retrieval Model for Geographic Information. In *Proceedings of the 2nd International Conference on Geographic Information Science (GIScience'02)*, pages 65–79, Boulder, CO, USA, September 25-28 2002. Springer.

Nuno Cardoso. Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas. Master's thesis, Faculty of Engineering, University of Porto, October 2006. In Portuguese.

Nuno Cardoso and Diana Santos. To Separate or not to Separate: Reflections About GIR Practice. In *1st Workshop on Novel Methodologies for Evaluation in Information Retrieval (NMEIR'08)*, Glasgow, UK, March 30 2008.

Nuno Cardoso and Mário J. Silva. Query Expansion through Geographical Feature Types. In *Proceedings of the 4th Workshop on Geographic Information Retrieval (GIR'07)*, Lisbon, Portugal, November 9 2007. ACM.

Nuno Cardoso, David Cruz, Marcirio Chaves, and Mário J. Silva. Using Geographic Signatures as Query and Document Scopes in Geographic IR. In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, Anselmo Peñas, Vivian Petras, and Diana Santos, editors, *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, volume 5152 of *LNCS*, pages 802–810. Springer, 2008.

Claudio Carpineto, Renato de Mori, Giovanni Romano, and Brigitte Bigi. An Information-Theoretic Approach to Automatic Query Expansion. *ACM Transactions on Information Systems*, 19(1):1–27, 2001.

Marcirio Chaves. *Uma Metodologia para Construção de Geo-Ontologias*. PhD thesis, University of Lisbon, Faculty of Sciences, December 2008. In Portuguese.

Marcirio Chaves and Diana Santos. What Kinds of Geographical Information Are There in the Portuguese Web? In Renata Vieira, Paulo Quaresma, Maria das Graças Volpe Nunes, Nuno J. Mamede, Claudia Oliveira, and Maria Carmelita Dias, editors, *Computational Processing of the Portuguese Language, 7th International Workshop, PROPOR 2006, Itatiaia, Brazil, May 13-17, 2006, Proceedings*, volume 3960 of *LNCS*, pages 264–267. Springer, 2006.

Marcirio Chaves, Mário J. Silva, and Bruno Martins. A Geographic Knowledge Base for Semantic Web Applications. In Carlos A. Heuser, editor, *20 Simpósio Brasileiro de Bancos de Dados (SBBD'2005)*, pages 40–54, Uberlândia, MG, Brazil, October 3-7 2005.

Marcirio Chaves, Catarina Rodrigues, and Mário J. Silva. Data Model for Geographic Ontologies Generation. In *XML: Aplicações e Tecnologias Associadas (XATA'2007)*, Lisboa, Portugal, February 15-16 2007.

Nancy Chinchor. The Statistical Significance of MUC-4 Results. In *Proceedings of the 4th Conference on Message Understanding, MUC-4*, pages 30–50. McLean, USA, June 16–18 1992.

Cyril W. Cleverdon. The Cranfield Tests on Index Language Devices. *Aslib Proceedings*, 19(6):173–193, 1967.

Paul Clough, Mark Sanderson, and Hideo Joho. Extraction of Semantic Annotations from Textual Web Pages. Technical Report Deliverable D15 6201, SPIRIT Project (EU IST-2001-35047), University of Sheffield, UK, 2004.

James R. Cowie and Wendy G. Lehnert. Information Extraction. *Communications of the ACM*, 39(1):80–91, 1996.

Steve Cronen-Townsend and W. Bruce Croft. Quantifying Query Ambiguity. In *Proceedings of the 2nd International Conference on Human Language Technology Research (HLT'02)*, pages 104–109, San Diego, CA, USA, 2002. Morgan Kaufmann.

Steve Cronen-Townsend, Yun Zhou, and Bruce Croft. Predicting Query Performance. In *Proceedings of the 25th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'02)*, pages 299–306, Tampere, Finland, August 11-15 2002. ACM.

Silviu Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'2007)*, Prague, Czech Republic, June 28-30 2007. ACL.

Hang Cui, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma. Probabilistic Query Expansion Using Query Logs. In *Proceedings of the 11th International Conference on World Wide Web (WWW'2002)*, pages 325–332, Honolulu, HI, USA, 2002. ACM.

Hamish Cunningham, Diana Maynard, and Valentin Tablan. JAPE: a Java Annotation Patterns Engine (Second Edition). Research Memorandum CS–00–10, Department of Computer Science, University of Sheffield, UK, November 2000.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, pages 168–175, Philadelphia, PA, USA, July 6-12 2002.

Tiago Delboni, Karla A. V. Borges, Alberto H. F. Laender, and Clodoveu A. Davis. Semantic Expansion of Geographic Web Queries Based on Natural Language Positioning Expressions. *Transactions in GIS*, 11(3):377–397, 2007.

Ian Densham and James Reid. A Geo-Coding Service Encompassing a Geo-Parsing Tool and Integrated Digital Gazetteer Service. In András Kornai and Beth Sundheim, editors,

Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, Edmonton, Canada, May 31 2003. ACL.

Junyan Ding, Luis Gravano, and Narayanan Shivakumar. Computing Geographical Scopes of Web Resources. In Amr El Abbadi, Michael L. Brodie, Sharma Chakravarthy, Umeshwar Dayal, Nabil Kamel, Gunter Schlageter, and Kyu-Young Whang, editors, *Proceedings of the 26th International Conference on Very Large Data Bases (VLDB'00)*, pages 545–556, Cairo, Egypt, September 10-14 2000. Morgan Kaufmann.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The Automatic Content Extraction (ACE) Program: Tasks, Data and Evaluation. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, pages 837–840, Lisboa, Portugal, May 26-28 2004. ELRA.

Aaron Douthat. The Message Understanding Conference Scoring Software User's Manual. In *Proceedings of the 7th Conference on Message Understanding, MUC-7*, Fairfax, USA, Abril 1998.

Efthimis N. Efthimiadis. A User-Centred Evaluation of Ranking Algorithms for Interactive Query Expansion. In Robert Korfhage, Edie M. Rasmussen, and Peter Willett, editors, *Proceedings of the 16th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'93)*, pages 146–159, Pittsburgh, PA, USA, June 27 - July 1 1993. ACM.

Efthimis N. Efthimiadis. Query Expansion. *Annual Review of Information Systems and Technology*, 31:121–187, 1996.

Max J. Egenhofer. Toward the Semantic Geospatial Web. In *Proceedings of the 10th ACM international symposium on Advances in Geographic Information Systems*, pages 1–4, McLean, Virginia, USA, November 8-9 2002.

Max J. Egenhofer and D. M. Mark. Naive geography. In A. U. Frank and W. Kuhn, editors, *Spatial Information Theory - A Theoretical Basis for GIS (COSIT'95)*, pages 1–15. Springer, Berlin, Heidelberg, 1995.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Unsupervised Named-Entity

Extraction from the Web: An Experimental Study. *Artificial Intelligence*, 165(1):91–134, 2005.

Christiane Fellbaum. *WordNet: an Electronic Lexical Database*. MIT Press, 1998.

Óscar Ferrández, Zornitza Kozareva, Andrés Montoyo, and Rafael Muñoz. NERUA: Sistema de Detección y Clasificación de Entidades Utilizando Aprendizaje Automático. *Procesamiento del Lenguaje Natural*, 35:37–44, 2005.

Regina Célia Figueiredo. Estudo comparativo de julgamentos de relevância do usuário e não usuário nos serviços de D.S.I. *Revista Ciência da Informação*, 7(2):69–78, 1978.

Larry Fitzpatrick and Mei Dent. Automatic Feedback Using Past Queries: Social Searching? In *Proceedings of the 20th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'97)*, pages 306–313, Philadelphia, PA, USA, July 27-31 1997. ACM.

Bruno M. Fonseca, Paulo Golgher, Bruno Pôssas, Berthier Ribeiro-Neto, and Nivio Ziviani. Concept-based Interactive Query Expansion. In Otthein Herzog, Hans-Jörg Schek, Norbert Fuhr, Abdur Chowdhury, and Wilfried Teiken, editors, *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM'05)*, pages 696–703, Bremen, Germany, October 31 – November 5 2005. ACM.

Gaihua Fu, Christopher B. Jones, and Alia I. Abdelmoty. Ontology-Based Spatial Query Expansion in Information Retrieval. In *On the Move to Meaningful Internet Systems: ODBASE 2005*, number 3761 in LNCS, pages 1466–1482. Springer, 2005.

Volker Gaede and Oliver Günther. Multidimensional Access Methods. *ACM Computing Surveys*, 30(2):170–231, 1998.

William A. Gale, Kenneth W. Church, and David Yarowsky. One Sense per Discourse. In *Proceedings of the Workshop on Speech and Natural Language (HLT'92)*, pages 233–237, Harriman, NY, USA, February 23-26 1992. ACL.

William A. Gale, Kenneth W. Church, and David Yarowsky. A Method for Disambiguating Word Senses in a Large Corpus. *Computers and the Humanities*, 26(5/6):415–439, 1993.

Qingqing Gan, Josh Attenberg, Alexander Markowetz, and Torsten Suel. Analysis of Geographic Queries in a Search Engine Log. In *Proceedings of the 1st International Workshop on Location and the Web (LOCWEB)'08*, pages 49–56, Beijing, China, 2008. ACM.

Frederic Gey, Ray Larson, Mark Sanderson, Hideo Joho, and Paul Clough. GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track. In Carol Peters, Frederic Gey, Julio Gonzalo, Henning Müeller, Gareth J.F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke, editors, *Acessing Multilingual information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF'2005. Revised Selected papers*, volume 4022 of *LNCS*, pages 908–919. Springer, 2006.

Fredric Gey, Ray Larson, Mark Sanderson, Kerstin Bischoff, Thomas Mandl, Christa Womser-Hacker, Diana Santos, Paulo Rocha, Giorgio Di Nunzio, and Nicola Ferro. Challenges to Evaluation of Multilingual Geographic Information Retrieval in GeoCLEF. In *Proceedings of the 1st International Workshop on Evaluating Information Access, EVIA 2007 (NTCIR-6 Pre-Meeting Workshop)*, Tokyo, Japan, May 15 2007a.

Fredric Gey, Ray Larson, Mark Sanderson, Kerstin Bishoff, Thomas Mandl, Christa Womser-Hacker, Diana Santos, Paulo Rocha, Giorgio Di Nunzio, and Nicola Ferro. GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. In Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, and Maximilian Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval: 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Revised selected papers*, volume 4730 of *LNCS*, pages 852–876. Springer, 2007b.

Luis Gravano, Vasileios Hatzivassiloglou, and Richard Lichtenstein. Categorizing Web Queries According to Geographical Locality. In *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM'03)*, pages 325–333, New Orleans, LA, USA, November 2-8 2003. ACM.

Daniel Gruhl, Laurent Chavet, David Gibson, Jörg Meyer, Pradhan Pattanayak, Andrew Tomkins, and Jason Y. Zien. How to Build a WebFountain: An Architecture for Very Large-Scale Text Analytics. *IBM Systems Journal*, 43(1):64–77, 2004.

David Haines and W. Bruce Croft. Relevance Feedback and Inference Networks. In *Proceedings of the 16th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'93)*, pages 2–11, Pittsburgh, PA, USA, June 27 - July 1 1993. ACM.

Donna Harman. Towards Interactive Query Expansion. In *Proceedings of the 11th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'88)*, pages 321–331, Grenoble, France, 1988. ACM.

Donna Harman. Relevance Feedback Revisited. In Nicholas J. Belkin, Peter Ingwersen, and Annelise Mark Pejtersen, editors, *Proceedings of the 15th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'92)*, pages 1–10, Copenhagen, Denmark, June 21-24 1992. ACM.

Patricia Harpring. Proper Words in Proper Places: The Thesaurus of Geographic Names. *MDA Information*, 3(2):5–12, 1997.

Stephen P. Harter. *A Probabilistic Approach to Automatic Keyword Indexing*. PhD thesis, University of Chicago, 1974.

Alexander G. Hauptmann and Andreas M. Olligschlaeger. Using Location Information from Speech Recognition of Television News Broadcasts. In Toby Robinson and Steve Renals, editors, *Proceedings of the ISCA (ESCA) Tutorial and Research Workshop on Accessing Information in Spoken Audio (Access-Audio-1999)*, pages 102–106, Cambridge, UK, April 19-20 1999. University of Cambridge.

Linda Ladd Hill. Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In José Luis Borbinha and Thomas Baker, editors, *Research and Advanced Technology for Digital Libraries, 4th European Conference, ECDL 2000, Lisbon, Portugal, September 18-20, 2000, Proceedings*, volume 1923 of *LNCS*, pages 280–290. Springer, 2000.

Linda Ladd Hill. *Georeferencing: The Geographic Associations of Information*. MIT Press, September 2006.

Linda Ladd Hill. *Access to Geographic Concepts in Online Bibliographic Files: Effectiveness of Current Practices and the Potential of a Graphic Interface*. PhD thesis, University of Pittsburgh, 1990.

Linda Ladd Hill, James Frew, and Qi Zheng. Geographic Names: The Implementation of a Gazetteer in a Georeferenced Digital Library. *D-Lib Magazine*, 1(5), 1999.

Lynette Hirschman. The Evolution of Evaluation: Lessons From the Message Understanding Conferences. *Computer Speech and Language*, 12(4):281–305, 1998.

E. Ide. New Experiments in Relevance Feedback. In Gerald Salton, editor, *The SMART Retrieval System – Experiments in Automatic Document Processing*, pages 337—354. Prentice-Hall, Englewood Cliffs, 1971.

ISO 19109. ISO 19109:2005 - Geographic Information - Rules for Application Schema, 2005. Available at `http://www.iso.org/`.

Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. *Information Processing and Management*, 36(2):207–227, 2000.

Chris Jones, Ross Purves, Anne Ruas, Mark Sanderson, Monika Sester, Marc J. van Kreveld, and Robert Weibel. Spatial Information Retrieval and Geographical Ontologies - An Overview of the SPIRIT project. In *Proceedings of the 25th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'02)*, pages 387–388, Tampere, Finland, August 11-15 2002.

Chris Jones, Alia I. Abdelmoty, David Finch, Gaihua Fu, and Subodh Vaid. The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. In Max J. Egenhofer, Christian Freksa, and Harvey J. Miller, editors, *Geographic Information Science, 3rd International Conference, GIScience 2004, Adelphi, MD, USA, October 20-23, 2004, Proceedings*, volume 3234 of *LNCS*, pages 125–139. Springer, 2004.

Christopher B. Jones, Harith Alani, and Douglas Tudhope. Geographical Information Retrieval with Ontologies of Place. In Daniel R. Montello, editor, *Spatial Information Theory: Foundations of Geographic Information Science, International Conference, COSIT 2001, Morro Bay, CA, USA, September 19-23, 2001, Proceedings*, volume 2205 of *LNCS*, pages 322–335. Springer, 2001.

Kalervo Järvelin, Jaana Kekäläinen, and Jaana Kekäläinen. The Impact of Query Structure and Query Expansion on Retrieval Performance. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*, pages 130–137, Melbourne, Australia, August 24–28 1998. ACM.

Graham Katz, Inderjeet Mani, and Thora Tenbrink, editors. *Workshop on Methodologies and Resources for Processing Spatial Language*, Marrakech, Morocco, May 31 2008. URL `http://www.lrec-conf.org/proceedings/lrec2008/workshops.html`.

Jun'ichi Kazama and Kentaro Torisawa. Exploiting Wikipedia as External Knowledge for Named Entity Recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'2007)*, pages 698–707, Prague, Czech Republic, June 28-30 2007. ACL.

Janet Kohler. Analysing Search Engine Queries for the Use of Geographic Terms. Master's thesis, University of Sheffield, 2003.

András Kornai. Evaluating Geographic Information Retrieval. In Carol Peters, Fredric C. Gey, Julio Gonzalo, Henning Müller, Gareth J. F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke, editors, *Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evalution Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers*, volume 4022 of *LNCS*, pages 928–938. Springer, 2006.

András Kornai and Beth Sundheim, editors. *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Edmonton, Canada, May 31 2003. ACL.

Reiner Kraft and Jason Zien. Mining Anchor Text for Query Refinement. In *Proceedings of the 13th International Conference on World Wide Web (WWW'2004)*, pages 666–674, New York City, NY, USA, May 17-22 2004.

Ray Larson. Geographic Information Retrieval and Spatial Browsing. In Linda C. Smith and Myke Gluck, editors, *Geographic information systems and libraries: patrons, maps, and spatial information: (papers presented at the 1995 Clinic on Library Applications of Data Processing, April 10-12, 1995)*, pages 81–124, 1996.

Ora Lassila and Ralph Swick. Resource Description Framework (RDF): Model and Syntax. W3C, World Wide Web Consortium, 1998. URL `http://www.w3.org/TR/WD-rdf-syntax/`.

Jochen Leidner. Toponym Resolution: A First Large-Scale Comparative Evaluation. Technical Report EDI-INF-RR-0839, School of Informatics, University of Edinburgh, July 2006.

Jochen Leidner. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. PhD thesis, University of Edinburgh, May 2007.

Huifeng Li, Rohini K. Srihari, Cheng Niu, and Wei Li. Infoxtract Location Normalization: a Hybrid Approach to Geographic References in Information Extraction. In András Kornai and Beth Sundheim, editors, *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, pages 39–44, Edmonton, Canada, May 31 2003. ACL.

Yi Li, Allistair Moffat, Nicola Stokes, and Lawrence Cavedon. Exploring Probabilistic Toponym Resolution for Geographical Information Retrieval. In Ross Purves and Chris Jones, editors, *Proceedings of the 3rd ACM Workshop On Geographic Information*

Retrieval, GIR 2006, Seattle, WA, USA, August 10, 2006, pages 17–22. Department of Geography, University of Zurich, 2006.

Zhisheng Li, Chong Wang, Xing Xie, and Wei-Ying Ma. Query Parsing Task for GeoCLEF 2007 Report. In Carol Peters et al., editors, *Working Notes of GeoCLEF 2007*, Budapest, Hungary, September 19-21 2007.

X. Allan Lu and Robert B. Keefer. Query Expansion/Reduction and its Impact on Information Retrieval Effectiveness. In Donna Harman, editor, *Proceedings of the 3rd Text REtrieval Conference (TREC'94)*, pages 231–239, Gaithersburg, MA, USA, 1995.

Hans Peter Luhn. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*, 1(4):309–317, October 1957.

Cristina López-Pujalte, Vicente P. Guerrero Bote, and Félix de Moya Anegón. A Test of Genetic Algorithms in Relevance Feedback. *Information Processing and Management*, 38(6):793–805, November 2002.

Bradley Malin. Unsupervised Name Disambiguation via Social Network Similarity. In *Workshop on Link Analysis, Counterterrorism, and Security in conjunction with the SIAM International Conference on Data Mining*, pages 93–102, Newport Beach, CA, USA, 2005.

Thomas Mandl, Paula Carvalho, Fredric Gey, Ray Larson, Diana Santos, and Christa Womser-Hacker. GeoCLEF 2008: the CLEF 2008 Cross-Language Geographic Information Retrieval Track Overview. In *Working Notes of CLEF 2008*, Aarhus, Denmark, September 17-19 2008a.

Thomas Mandl, Fredric Gey, Giorgio Di Nunzio, Nicola Ferro, Ray Larson, Mark Sanderson, Diana Santos, Christa Womser-Hacker, and Xing Xie. GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview. In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, Anselmo Peñas, Vivian Petras, and Diana Santos, editors, *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, volume 5251 of *LNCS*. Springer, 2008b.

Alexander Markowetz, Yen-Yu Chen, Torsten Suel, Xiaohui Long, and Bernhard Seeger. Design and Implementation of a Geographic Search Engine. In AnHai Doan, Frank Neven, Robert McCann, and Geert Jan Bex, editors, *Proceedings of the*

8th International Workshop on the Web & Databases (WebDB'2005), pages 19–24, Baltimore, MA, USA, June 16-17 2005.

Bruno Martins. *Geographically Aware Web Text Mining*. PhD thesis, University of Lisbon, Faculty of Sciences, August 2008.

Bruno Martins and Mário J Silva. A Graph-Ranking Algorithm for Geo-Referencing Documents. In *Proceeedings of the 5th IEEE International Conference on Data Mining (ICDM'05)*, pages 741–744, Houston, TX, USA, November 27–30 2005. IEEE Computer Society.

Bruno Martins, Mário J. Silva, and Leonardo Andrade. Indexing and Ranking in Geo-IR Systems. In Otthein Herzog, Hans-Jörg Schek, Norbert Fuhr, Abdur Chowdhury, and Wilfried Teiken, editors, *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM'05)*, pages 31–34, Bremen, Germany, October 31 – November 5 2005. ACM.

Bruno Martins, Nuno Cardoso, Marcirio Chaves, Leon ardo Andrade, and Mário J. Silva. The University of Lisbon at GeoCLEF 2006. In Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke, and Maximilian Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Alicante, Spain, September, 2006. Revised Selected papers*, volume 4730 of *LNCS*, pages 986–994. Springer, Setember 2007.

Kevin S. McCurley. Geospatial Mapping and Navigation of the Web. In *Proceedings of the 10th International Conference on World Wide Web (WWW'01)*, pages 221–229, Hong Kong, China, May 1-5 2001. ACM.

David D. McDonald. Internal and External Evidence in the Identification and Semantic Categorization of Proper Names. In Branimir Boguraev and James Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*, pages 61–76. MIT Press, 1993.

Andrei Mikheev, Marc Moens, and Claire Grover. Named Entity Recognition without Gazetteers. In *Proceedings of the 9th European Chapter of the Association for Computational Linguistics (EACL'99)*, pages 1–8, Bergen, Norway, June 8-12 1999. ACL.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244, 1990.

Mandar Mitra, Amit Singhal, and Chris Buckley. Improving Automatic Query Expansion. In *Proceedings of the 21st Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'98)*, pages 206–214, Melbourne, Australia, August 24-28 1998. ACM Press.

Stefano Mizzaro. Relevance: The Whole History. *Journal of the American Society for Information Science*, 48(9):810–832, 1997.

Cristina Mota and Max Silberztein. Em busca da máxima precisão sem almanaques. O Stencil/Nooj no HAREM. In Diana Santos and Nuno Cardoso, editors, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, chapter 15, pages 191–208. Linguateca, 2007. In Portuguese.

David Nadeau. Balie – Baseline Information Extraction: Multilingual Information Extraction from Text with Machine Learning and Natural Language Techniques. Technical report, University of Ottawa, 2005.

David Nadeau. *Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision*. PhD thesis, University of Ottawa, 2007.

Simon E. Overell. *Geographic Information Retrieval: Classification, Disambiguation and Modelling*. PhD thesis, Department of Computing, Imperial College, 2009.

Simon E. Overell and Stefan M. Rüger. Identifying and Grounding Descriptions of Places. In Ross Purves and Chris Jones, editors, *Proceedings of the 3rd ACM Workshop On Geographic Information Retrieval, GIR 2006, Seattle, WA, USA, August 10, 2006*. Department of Geography, University of Zurich, 2006.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

David D. Palmer and David S. Day. A Statistical Profile of the Named Entity Task. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP'97)*, pages 190–193, Washington, DC, USA, 1997. ACL.

Greg Pass, Abdur Chowdhury, and Cayley Torgeson. A picture of search. In *Proceedings of the 1st International Conference on Scalable Information Systems*, Hong Kong, June 2006.

Carol Peters and Martin Braschler. Cross-Language System Evaluation: the CLEF campaigns. *Journal of the American Society of Information Science*, 52(12):1067–1072, 2001.

Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, and Tom De Groeve. Geographical Information Recognition and Visualization in Texts Written in Various Languages. In *Proceedings of the 2004 ACM Symposium on Applied Computing (SAC'04)*, pages 1051–1058, Nicosia, Cyprus, 2004. ACM.

David V. Pullar and Max J. Egenhofer. Toward Formal Definitions of Topological Relations Among Spatial Objects. In Duane Marble, editor, *3rd International Symposium on Spatial Data Handling (SDH'88)*, pages 225–441, Sydney, Australia, August 1988.

Ross Purves and Chris Jones. Workshop on Geographic Information Retrieval. *Computers, Environment and Urban Systems*, 30(4):375–377, 2006.

Yonggang Qiu and Hans Peter Frei. Concept Based Query Expansion. In *Proceedings of the 16th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'93)*, pages 160–169, Pittsburgh, PA, USA, June 27 - July 1 1993. ACM.

Erik Rauch, Michael Bukatin, and Kenneth Baker. A Confidence-Based Framework for Disambiguating Geographic Terms. In András Kornai and Beth Sundheim, editors, *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, pages 50–54, Edmonton, Canada, May 31 2003. ACL.

James Reid. geoXwalk - A Gazetteer Server and Service for UK Academia. In Traugott Koch and Ingeborg Sølvberg, editors, *Research and Advanced Technology for Digital Libraries, 7th European Conference, ECDL 2003, Trondheim, Norway, August 17-22, 2003, Proceedings*, volume 2769 of *LNCS*, pages 387–392. Springer, 2003.

Stephen E. Robertson. On Relevance Weight Estimation and Query Expansion. *Journal of Documentation*, 42(3):182–188, 1986.

Stephen E. Robertson. On Term Selection for Query Expansion. *Journal of Documentation*, 46(4):359–364, 1990.

Stephen E. Robertson and Karen Spärck Jones. Simple Proven Approaches to Text Retrieval. Technical report, University of Cambridge, May 1997.

Stephen E. Robertson and Karen Spärck Jones. Relevance Weighting of Search Terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.

Stephen E. Robertson and Steven G. Walker. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In W. Bruce Croft and C. J. van Rijsbergen, editors, *Proceedings of the 17th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'94)*, pages 232–241, Dublin, Ireland, July 3-6 1994. ACM/Springer.

Stephen E Robertson, Steven G. Walker, Micheline Hancock-Beaulieu, Aarron Gull, and Marianna Lau. Okapi at TREC. In *Proceedings of the 1st Text REtrieval Conference (TREC'92)*, pages 21–30. National Institute of Standards and Technology (NIST), 1992. Special Publication 500-207.

Stephen E. Robertson, Steven G. Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In *Proceddings of the 3rd Text REtrieval Conference (TREC'94)*. NIST, November 1994.

J. J. Rocchio Jr. Relevance Feedback in Information Retrieval. In Gerard Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice-Hall, Englewood Cliffs, 1971.

Catarina Rodrigues. An Ontology of the Physical Geography of Portugal. Master's thesis, University of Lisbon, Faculty of Sciences, October 2008.

Ian Ruthven and Mounia Lalmas. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review*, 18(2):95–145, 2003.

Gerald Salton. *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice Hall, NJ, USA, 1971.

Gerald Salton and Chris Buckley. On the Use of Spreading Activation Methods in Automatic Information. In *Proceedings of the 11th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'88)*, pages 147–160, Grenoble, France, 1988. ACM.

Gerald Salton and Chris Buckley. Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for (I)nformation Science*, 41(4):288–297, 1990.

Gerald Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620, November 1975.

Mark Sanderson and Bruce Croft. Deriving Concept Hierarchies from Text. In *Proceedings of the 22nd Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'99)*, pages 206–213, Berkeley, CA, USA, 1999. ACM.

Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2002*, pages 155–158, Taipei, Taiwan, 2002.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2003*, pages 142–147, Edmonton, Canada, 2003.

Diana Santos. *Translation-based corpus studies: Contrasting Portuguese and English tense and aspect systems*. Rodopi, Amsterdam/New York, NY, USA, 2004.

Diana Santos. Evaluation in Natural Language Processing. In *Foundational course, European Summer School on Language, Logic and Information, ESSLLI 2007*, Dublin, Ireland, August 6-10 2007. URL `http://www.linguateca.pt/Diana/download/SantosESSLLI2007rev.pdf`.

Diana Santos and Nuno Cardoso. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, 2007.

Diana Santos and Marcirio Chaves. The Place of Place in Geographical IR. In Ross Purves and Chris Jones, editors, *Proceedings of the 3rd ACM Workshop On Geographic Information Retrieval, GIR 2006, Seattle, WA, USA, August 10, 2006*. Department of Geography, University of Zurich, 2006.

Diana Santos, Paula Carvalho, Hugo Oliveira, and Cláudia Freitas. Second HAREM: new challenges and old wisdom. In António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira, and Paulo Quaresma, editors, *Computational Processing of Portuguese Language, 8th International Conference (PROPOR'2008), September 8-10, Aveiro, Portugal. Proceedings*, number 5190 in LNCS, pages 212–215. Springer, 2008.

Diana Santos, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, Johannes Leveling, and Yvonne Skalban. GikiP at GeoCLEF 2008: Joining GIR and QA forces for querying Wikipedia. In Carol Peters, Tomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F.Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, and Viviane Petras, editors, *Evaluating Systems for Multilingual and Multimodal Information Access: 9th*

Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers. Springer, 2009.

Frank Schilder, Yannick Versley, and Christopher Habel. Extracting spatial information: grounding, classifying and linking spatial expressions. In *Workshop on Geographic Information Retrieval*, Sheffield, UK, 2004. ACM.

Satoshi Sekine. On-demand Information Extraction. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*. ACL, 2006.

Benny Shanon. Where Questions. In *Proceedings of the 17th Annual Meeting on Association for Computational Linguistics*, pages 73–75, La Jolla, CA, USA, 29 June – 1 July 1979. ACL.

Yusuke Shinyama and Satoshi Sekine. Preemptive Information Extraction using Unrestricted Relation Discovery. In Robert C. Moore, Jeff A. Bilmes, Jennifer Chu-Carroll, and Mark Sanderson, editors, *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 4-9, 2006, New York, NY, USA*. ACL, 2006.

Max Silberztein. NooJ: A Cooperative, Object-Oriented Architecture for NLP. *INTEX pour la Linguistique et le traitement automatique des langues*, 2004. Cahiers de la MSH Ledoux.

Mário J. Silva, Bruno Martins, Marcirio Chaves, Ana Paula Afonso, and Nuno Cardoso. Adding Geographic Scopes to Web Resources. *CEUS - Computers Enviroment and Urban Systems*, 30(4):378–399, 2006.

Nuno Silva. Pesquisa de Informação e a Web Semântica. In *3ª TeIA - Tertúlias em Inteligência Artificial*, Porto, Portugal, October 17 2007. ISEP. Unpublished presentation.

Amit Singhal. Modern Information Retrieval: A Brief Overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 24(4):35–43, 2001.

Amit Singhal. Web Search: Challenges and Directions. In Craig MacDonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White, editors, *Advances in Information Retrieval, 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings*, volume 4956 of *LNCS*. Springer, 2008.

Alan F. Smeaton. Information Retrieval: Still Butting Heads with Natural Language Processing? In Maria Teresa Pazienza, editor, *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, International Summer School, SCIE-97, Frascati, Italy, 14-18, 1997*, volume 1299 of *LNCS*, pages 115–138. Springer, 1997.

Barry Smith and David M. Mark. Geographical Categories: an Ontological Investigation. *International Journal of Geographical Information Science*, 15(7):591–612, 2001.

David A. Smith and Gregory Crane. Disambiguating Geographic Names in a historical Digital Library. In Panos Constantopoulos and Ingeborg Sølvberg, editors, *Research and Advanced Technology for Digital Libraries, 5th European Conference, ECDL 2001, Darmstadt, Germany, September 4-9, 2001,Proceedings*, volume 2163 of *LNCS*, pages 127–136. Springer, 2001.

Michael K. Smith, Chris Welty, and Deborah L. McGuinness. OWL Web Ontology Language Guide. W3C, World Wide Web Consortium, 2004. URL `http://www.w3.org/TR/owl-guide/`.

Karen Spärck Jones. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28:11–21, 1972.

Karen Spärck Jones. IDF Term Weighting and IR Research Lessons. *Journal of Documentation*, 60(5):521–523, 2004.

Karen Spärck Jones. *Automatic Keyword Classification for Information Retrieval*. Butterworths, London, UK, 1971.

Nicola Stokes, Yi Li, Alistair Moffat, and Jiawen Rong. An Empirical Study of the Effects of NLP Components on Geographic IR Performance. *International Journal of Geographical Information Science*, 22(3):247–264, 2008.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO: A Core of Semantic Knowledge. In Carey L. Williamson, Mary Ellen Zurko, Peter F. Patel-Schneider, and Prashant J. Shenoy, editors, *Proceedings of the 16th International World Wide Web conference (WWW'07)*, Banff, Canada, May 8-12 2007. ACM.

Beth Sundheim. Overview of Results of the MUC-6 Evaluation. In *Proceedings of the 6th Conference on Message Understanding, MUC-6*, pages 13–31, Columbia, SC, USA, November 6-8 1995.

Howard Turtle. *Inference Networks for Document Retrieval*. PhD thesis, Department of Computer Science, University of Massachusetts, Amherst, 1990.

Subodh Vaid, Christopher B. Jones, Hideo Joho, and Mark Sanderson. Spatio-textual Indexing for Geographical Search on the Web. In Claudia Bauzer Medeiros, Max J. Egenhofer, and Elisa Bertino, editors, *Advances in Spatial and Temporal Databases, 9th International Symposium, SSTD 2005, Angra dos Reis, Brazil, August 22-24, 2005, Proceedings*, volume 3633 of *LNCS*, pages 218–235. Springer, 2005.

Marc van Kreveld, Iris Reinbacher, Avi Arampatzis, and Roelof van Zwol. Distributed Ranking Methods for Geographic Information Retrieval. In Peter F. Fisher, editor, *Proceedings of the 11th International Symposium on Spatial Data Handling: Developments in Spatial Data Handling*, pages 231–243. Springer, August 23-25 2004.

C. J. van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, 1979. 2nd edition.

Ellen M. Voorhees. Query Expansion using Lexical-Semantic Relations. In W. Bruce Croft and C. J. van Rijsbergen, editors, *Proceedings of the 17th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'94)*, pages 61–69, Dublin, Ireland, July 3-6 1994. ACM/Springer.

Piek Vossen. *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA, 1998.

Max Völkel, Markus Krötzsch, Denny Vrandecic, Heiko Haller, and Rudi Studer. Semantic wikipedia. In *Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland, UK, May 23-26, 2006*, pages 585–594. ACM, 2006.

Nina Wacholder, Yael Ravin, and Misook Choi. Disambiguation of Proper Names in Text. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP'97)*, pages 202–208, Washington, DC, USA, March 31 – April 3 1997. ACL.

Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005. 2nd Edition.

Ian H. Witten, Alistair Moffat, and Timothy C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann, 1994.

Allison Woodruff and Christian Plaunt. GIPSY: Automated Geographic Indexing of Text Documents. *Journal of the American Society for Information Science*, 45(9):645–655, 1994.

Fei Wu and Daniel S. Weld. Autonomously Semantifying Wikipedia. *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM'07)*, pages 41–50, November 6-8 2007.

Jinxi Xu and W. Bruce Croft. Improving the Effectiveness of Information Retrieval with Local Context Analysis. *ACM Transactions on Information Systems*, 18(1):79–112, 2000.

Jinxi Xu and W. Bruce Croft. Query Expansion Using Local and Global Document Analysis. In Hans-Peter Frei, Donna Harman, Peter Schäuble, and Ross Wilkinson, editors, *Proceedings of the 19th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR'96)*, pages 4–11, Zurich, Switzerland, August 18-22 1996.

Jing-Jye Yang and Robert Korfhage. Query Modification Using Genetic Algorithms in Vector Space Models. *International Journal of Expert Systems*, 7(2):165–191, 1994.

Yinghua Zhou, Xing Xie, Chuang Wang, Yuchang Gong, and Wei-Ying Ma. Hybrid Index Structures for Location-based Web Search. In Otthein Herzog, Hans-Jörg Schek, Norbert Fuhr, Abdur Chowdhury, and Wilfried Teiken, editors, *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM'05)*, pages 155–162, Bremen, Germany, October 31 – November 5 2005. ACM.

Wenbo Zong, Dan Wu, Aixin Sun, Ee-Peng Lim, and Dion Hoe-Lian Goh. On Assigning Place Names to Geography Related Web Pages. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'05)*, pages 354–362. IEEE, June 7-11 2005.

Óscar Ferrández, Zornitsa Kozareva, Antonio Toral, Elisa Noguera, Andrés Montoyo, Rafael Muñoz, and Fernando Llopis. University of Alicante at GeoCLEF 2005. In Carol Peters, Frederic Gey, Julio Gonzalo, Henning Müeller, Gareth J.F. Jones, Michael Kluck, Bernardo Magnini, and Maarten de Rijke, editors, *Acessing Multilingual information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF'2005. Revised Selected papers*, volume 4022 of *LNCS*, pages 924–927. Springer, 2006.

Contents